



PHD

Bayesian Decision Making in Adaptive Clinical Trials

Burnett, Thomas

Award date:
2017

Awarding institution:
University of Bath

[Link to publication](#)

Alternative formats

If you require this document in an alternative format, please contact:
openaccess@bath.ac.uk

Copyright of this thesis rests with the author. Access is subject to the above licence, if given. If no licence is specified above, original content in this thesis is licensed under the terms of the Creative Commons Attribution-NonCommercial 4.0 International (CC BY-NC-ND 4.0) Licence (<https://creativecommons.org/licenses/by-nc-nd/4.0/>). Any third-party copyright material present remains the property of its respective owner(s) and is licensed under its existing terms.

Take down policy

If you consider content within Bath's Research Portal to be in breach of UK law, please contact: openaccess@bath.ac.uk with the details. Your claim will be investigated and, where appropriate, the item will be removed from public view as soon as possible.

Bayesian Decision Making in Adaptive Clinical Trials

submitted by

Thomas Burnett

for the degree of Doctor of Philosophy

of the

University of Bath

Department of Mathematical Sciences

July 23, 2017

COPYRIGHT

Attention is drawn to the fact that copyright of this thesis rests with its author. This copy of the thesis has been supplied on the condition that anyone who consults it is understood to recognise that its copyright rests with its author and that no quotation from the thesis and no information derived from it may be published without the prior written consent of the author.

This thesis may be made available for consultation within the University Library and may be photocopied or lent to other libraries for the purposes of consultation.

Signature of Author

Thomas Burnett

Summary

The key original contribution of this work is the use of a Bayes optimisation framework for the decision made at the interim analysis of Adaptive Enrichment trials.

Adaptive Enrichment designs make efficient use of pre-identified patient sub-populations. They begin by recruiting from all eligible patients, then at a pre-planned interim analysis select which sub-populations will be recruited from for the remainder of the sample. We ensure strong control of the Familywise Error Rate whichever sub-populations are selected by constructing an overall hypothesis testing structure using both closed testing procedures and combination tests. This allows us to make interim decision by any method we choose. We find the Bayes optimal decision, recruiting the remainder of the trial to optimise the Bayes expected gain of the trial. We compare the Bayes optimal Adaptive Enrichment trials with fixed sampling designs to understand the overall advantage of using adaptive trials.

This optimisation framework is very flexible, we evaluate the performance of Bayes optimal Adaptive Enrichment designs for different forms of data: delayed responses, longitudinal analysis and discuss the extension of these methods to survival data. Through this we see that although the information at the interim analysis is reduced the adaptive trials still offer some benefit. Additionally we investigate what may happen when we alter the pattern of recruitment of the Adaptive Enrichment trials, showing that adaptation may be useful in a broad range of scenarios.

Contents

1	Introduction	6
1.1	Adaptive Enrichment trials	6
1.2	Roadmap of this work	10
2	Background clinical trial methodology	12
2.1	A simple clinical trial	12
2.2	Conducting a simple trial	14
3	Conducting Adaptive Enrichment trials	19
3.1	Introducing sub-populations	19
3.1.1	Identification of sub-populations	19
3.1.2	Formulating the problem	20
3.2	Testing multiple hypotheses	22
3.2.1	Familywise error rate	22
3.2.2	Closed testing procedures	24
3.2.3	Procedures that give strong control of the FWER are closed testing procedures	25
3.2.4	Consonance	27
3.3	Fixed sampling designs	27
3.3.1	Formulation 1: single sub-population	27
3.3.2	Formulation 2: double sub-population	29
3.3.3	Simes method for testing the intersection hypothesis . . .	30
3.3.4	Example: Formulation 2	33
3.3.5	Example: Formulation 1	35
3.3.6	Fixed enrichment	37
3.4	Combination tests	38

3.4.1	Weighted inverse normal	40
3.5	Conducting adaptive enrichment trials	42
3.5.1	Adaptive enrichment methodology	42
3.5.2	Analysis	46
3.5.3	Example	47
3.6	An alternative presentation of the summary statistics	48
4	Evaluating performance and optimisation	53
4.1	Methods for comparison	53
4.1.1	Running example	53
4.1.2	Formulation 1	54
4.1.3	Formulation 2	57
4.2	Constructing a single measure of trial performance	60
4.2.1	Weighting outcomes	60
4.2.2	Gain functions	63
4.3	Optimisation of adaptive enrichment trials	74
4.3.1	Oracle decision rule	74
4.3.2	Plotting optimal decision boundaries for Adaptive Enrichment trials	78
4.3.3	Incorporating prior uncertainty	85
4.3.4	Simple decision optimisation	89
4.4	Bayes optimal decision	92
4.4.1	Comparing the Bayes optimal rule with the optimised simple decision rules	96
4.5	Conclusions	97
5	Delayed response and survival endpoints	98
5.1	Adaptive Enrichment with delayed response	98
5.1.1	Recruitment	98
5.1.2	Hypothesis testing	100
5.1.3	Optimisation	101
5.1.4	Example	102
5.2	Longitudinal analysis	107
5.2.1	Data collection	107
5.2.2	Interim optimisation	110

5.2.3	Adding to the joint model	116
5.2.4	Example	116
5.3	Survival endpoints	118
5.3.1	A single survival endpoint	118
5.3.2	Using progression free survival to improve decision making at the interim analysis	122
6	Sensitivity of optimal designs to the model assumptions	124
6.1	Changing the prior distribution	124
6.1.1	Returning to immediate response	124
6.1.2	Alternative prior distributions	125
6.1.3	Uniform prior distributions	127
6.1.4	Comparing an alternative prior distribution	128
6.2	Changing the parameters of the design	131
6.2.1	Sub-population proportion	131
6.2.2	Timing of the interim analysis	133
6.2.3	Total sample size of the trial	134
7	Variations to the Adaptive Enrichment design	136
7.1	Hypothesis testing methodology	136
7.1.1	Improved Simes method for the intersection hypothesis . .	137
7.1.2	Dunnet type rule for the intersection hypothesis	139
7.1.3	χ^2 combination test	142
7.1.4	Changing the ordering of testing methods	144
7.2	Early stopping for futility	146
8	Conclusions and some possible extensions	150
8.1	Conclusions	150
8.2	Extensions	152
8.2.1	Parameters governing recruitment	152
8.2.2	Sample size	153
8.2.3	Niche population	153
8.2.4	Similar and simplified decisions	154
8.2.5	Overall treatment effect estimates	154
8.2.6	Multiple interim analyses	154

A	A summary of techniques for practical use	156
A.1	Recruitment of trials	156
A.2	Testing methodologies	159
A.2.1	FamilyWise Error Rate	159
A.2.2	Closed testing procedure	159
A.2.3	Combination tests	161
A.2.4	Overall testing procedure	162
A.3	Defining the gain function	163
A.4	Optimisation of adaptive enrichment	164
A.4.1	Assessing overall trial performance	164
A.4.2	Optimising simple decision rules	164
A.4.3	Bayes optimal decisions	165
A.4.4	Early stopping for futility	166
A.5	When is Adaptive Enrichment useful	167
A.5.1	Oracle Adaptive Enrichment	167
A.6	Additional data types	168
A.6.1	Delayed response and survival endpoints	168
A.6.2	Longitudinal data	169

Chapter 1

Introduction

1.1 Adaptive Enrichment trials

The aim of this work is to use a Bayesian decision framework to optimise design choices at the interim analysis in adaptive clinical trials. We focus on the optimisation of the decision at the interim analysis of Adaptive Enrichment trials.

When the patient population for a clinical trial may be split into patient sub-populations we would like to recruit only the patients that receive a benefit from the new treatment. This decision may be difficult to make before beginning the trial; Adaptive Enrichment trials delay the selection of sub-populations until an interim analysis where the decision can be made based on trial observations. For example if we have two sub-populations we may recruit the first half of the sample from both, then at an interim analysis we may choose whether to recruit the remaining sample from either sub-population only or continue recruiting from both. In addition to optimising the decision at the interim analysis we may also use the Bayesian decision framework to find where using an Adaptive Enrichment trial is beneficial compared to a fixed sampling design.

The trials that we consider in this work are intended for use in Phase III of the clinical development process. We will work in the setting of randomised controlled trials, where patients are randomised to the new treatment and a control in order to establish whether the new treatment offers some improvement. This is the

same setting as Turnbull and Jennison (2000) consider for their group sequential methods; group sequential designs use multiple interim analyses throughout the trial to allow early stopping reducing the expected sample size and are well established methods for conducting clinical trials.

The idea of Adaptive Enrichment designs stem from the Adaptive Signature design proposed by Freidlin and Simon (2005), these Adaptive Signature trials differ from Adaptive Enrichment in that they attempt to identify the patient sub-populations during the trial whereas Adaptive Enrichment trials make use of pre specified sub-populations. Temple (2005) discusses the potential improvement in detecting efficacy of a new treatment by making use of sub-populations, where this improvement would come from selecting the appropriate sub-population. Bhatt and Mehta (2016) present some recent examples of adaptive trials and discuss the use of Adaptive Enrichment in this context, discussing the potential benefits of more widespread use of adaptive trials such as accelerating the development process. Bhatt and Mehta (2016) also discuss some regulatory concerns for adaptive trials but state that “to date, regulatory agencies have opined favourably about adaptive designs”.

Regulatory bodies have published their thoughts on the use of adaptive clinical trials in a confirmatory setting. Elsässer et al. (2014) give a summary of scientific advice letters from the European Medicines Agency for adaptive clinical trials. The FDA published guidelines for adaptive clinical trials Food and Drug Administration and others (2010), from the discussion in these guidelines the Adaptive Enrichment designs that fall under the “less well-understood” methods. The guidance in this setting would be useful to anyone considering the use of an adaptive design for a clinical trial. The key consideration we make from the guidelines in our work is the need to control the study wide type I error rate. Controlling the error rate where multiplicity is present in the design is a well established area, points to consider are given by Committee for Proprietary Medicinal Products and others (2002) and guidance is included in the ICH E9 document International Conference on Harmonisation E9 Expert Working Group and others (1999). Dmitrienko et al. (2013) cover key issues of multiplicity in the context of clinical trials.

The papers by Bretz et al. (2006) and Schmidli et al. (2006) discuss the selection of hypotheses at an interim analyses in a similar context to which we are now working. The analyses of these designs require some additional techniques since there are multiple hypotheses and due to the interim observations of the data, Bauer and Kohne (1994) discuss appropriate methods for experiments where an adaptive interim analysis is to be used. For the analysis of multiple hypotheses, a closed testing procedures Marcus et al. (1976) are used. For analysis over multiple stages we use combination tests as suggested by Bauer and Kohne (1994), to combine the stages of the trial. Jennison and Turnbull (2007) provide further discussion of the appropriate methods of hypothesis testing for Adaptive Enrichment trials. Simulation studies by Wang et al. (2009) have shown how the error rate is well controlled by these techniques, although we will use methods where this may be proven directly.

Rosenblum (2014) examine the overall performance of Adaptive Enrichment trials in terms of the power to reject null hypotheses, they show that any single fixed sampling design cannot dominate a comparable Adaptive Enrichment trial in terms of rejecting null hypotheses. This result may seem to be an unsurprising property but highlights two things: firstly it shows that the adaptive design is doing something unique when compared with fixed sampling alternatives, suggesting that the adaptation may be useful in some way; secondly we note that there are multiple fixed sampling alternatives that we should consider comparing our optimal Adaptive Enrichment trials with in order to make a complete assessment of their overall benefit.

We use a Bayesian optimisation framework for the interim decision of the Adaptive Enrichment trials, the book by Berger (2013) offers a broad introduction to statistical decision making with Bayesian methods. We construct a gain function as an overall measure of trial performance and seek to make the decision at the interim analysis that maximises this. The use of Bayesian methods has also been addressed by regulatory bodies for example the FDA guidance US Food and Drug Administration and others (2010) in the context of medical device trials. The main concern here is demonstrating that these methods do not inflate the error rate, we ensure this is the case with the methods we use for hypothesis

testing.

The Bayes optimisation allows us to construct Bayes optimal Adaptive Enrichment trials across a range of scenarios. We find that these optimal adaptive designs offer an improvement of the overall performance of the trial when compared to fixed sampling alternatives. Our optimisation is sufficiently flexible that it is not limited to the particular scenarios that we present, for example we are able to easily incorporate other possible interim decisions such as early stopping. Posch and Bauer (2013) consider Adaptive budgets in clinical trials, their work introduces a cost of sampling into the utility function. We use this cost of sampling to motivate early stopping of the entire trial for futility in our utility function; this could be further extended to allow for an adaptive sample size but we have not done so within this work.

Our optimisation is also not restricted to the type of data being used, one of the examples of alternative data types we have included is the use of survival endpoints. Schmidli et al. (2007) looked at the use of Bayesian methods in the interim analysis of seamless phase II/III trials with survival endpoints, both Jenkins et al. (2011) and Irle and Schäfer (2012) use the log rank score statistic to define the appropriate distributions for a trial using survival endpoints. We take the same approach as Jenkins et al. (2011) simulating trials using only the log rank score statistics. Another alternative is to consider longitudinal observations, in this case we follow the method of Hampson and Jennison (2013) to decrease the variance of the estimate of the final treatment effect at the time of the interim analysis.

In the literature there have not been any such extensive approaches to the optimisation of the interim decision in Adaptive Enrichment trials. Brannath et al. (2009) use Bayesian decision tools to optimise the decision at the interim analysis, however their focus is the Bayesian predictive probability of rejecting any null hypothesis. Brannath et al. (2009) do not explicitly define a utility function in their optimisation, if we construct the equivalent utility function we see that this is a special case of our form of optimisation; this is computationally simpler as finding the interim decision depends only on the posterior predictive

distribution. Götte et al. (2015) go about the optimisation of the interim decision quite differently, where they seek to maximise the probability of a correct decision, as with the optimisation of Brannath et al. (2009) this could also be written as a special case of our version of optimisation under the correct choice of utility function.

The Bayesian decision framework we use to optimise the decision at the interim analysis of Adaptive Enrichment trials may also be used to compare the overall trial performance with fixed sampling alternatives. The paper by Ondra et al. (2016) recently took this approach to comparing competing design choices (including Adaptive Enrichment) for a particular trial, they did not optimise the interim decision of the trial. Thus our use of the Bayesian decision framework to optimise Adaptive Enrichment trials extends upon the current literature in this aspect. We are also able to compare the overall performance of the Bayes optimal Adaptive Enrichment trials to show the maximum benefit that may be achieved by using an Adaptive Enrichment trial in any particular scenario.

1.2 Roadmap of this work

In Chapter 2 we introduce randomised controlled trials in a simple form, defining our notation and discussing the basic principles of conducting a trial. Chapter 3 extends immediately from this, introducing the two formulations of the problem that we consider when conducting Adaptive Enrichment trials. In this Chapter we introduce all of the necessary hypothesis testing methods and describe the recruitment of Adaptive Enrichment trials and the fixed sampling alternatives used throughout the work.

In Chapter 4 we evaluate the overall performance of Adaptive Enrichment trials in comparison to the fixed sampling alternatives using a gain function as a single measure. We introduce Oracle decision rules as the first form of optimisation, this is the best possible optimisation as it assumes the true treatment effects are known. We then introduce the Bayes decision framework allowing us to optimise rules of a simple form and ultimately find the Bayes optimal decisions. Using the same framework we compare the performance advantage of these optimised

decision rules, learning about the improvement the optimal adaptive designs may offer over fixed sampling trials.

In Chapter 5 we remove the simplifying assumption that observations are immediately available, applying our decision making framework to more complicated data. First we look at a delayed response, in this setting we have less information to make our decision but still have the same remaining sample size to alter. Despite these extra complications we are still able to use the same optimisation framework to optimise the interim decision and to evaluate the overall performance. Using survival endpoints we find that using the log-rank score statistic to summarise the trial gives the equivalent performance to that seen when assuming a delay to responses. We then see that if we have longitudinal observations the interim decisions may be enhanced by assuming a joint model for the observations over time, this recovers some of the performance lost in the designs assuming a delayed response.

In Chapter 6 we examine the sensitivity of the Adaptive Enrichment trial to the values of the design parameters we chose in earlier Chapters, we vary the prior distributions and the recruitment patterns and evaluate the effect on the overall performance of the designs. In addition to these parameters we also made choices about how any hypothesis tests would be conducted and what decisions would be available at the interim analysis, we examine how the overall performance may be improved by changing the hypothesis testing methods or allowing for early stopping for futility in Chapter 7.

In Chapter 8 we summarise our conclusions from our work and discuss some possible further extensions to the methods that we have presented.

Chapter 2

Background clinical trial methodology

2.1 A simple clinical trial

Randomised controlled clinical trials are used to establish the efficacy and safety of new treatments in comparison to the current standard of care, statistical principles for clinical trials are given by ICH Steering Committee and others (1998). We begin by discussing a randomised controlled trial in their simplest form. This allows the introduction of necessary notation and explanation of assumptions that have been made. The purpose of a randomised controlled trial is to find out whether a new treatment is more effective than a control treatment for treating some illness. The control is commonly an existing treatment that is the current standard of care for the illness, or where this is not available a suitable placebo is used. The trial will provide evidence about the effect of the new treatment in comparison to the control. Typically the trial will be conducted in some sub-set of the population with the illness, based on specific eligibility criteria. The trial designs presented in this work are intended to be suitable for use as Phase III or confirmatory trials, and so hypothesis testing will be required to ensure control of the type I error rate.

Suppose we have a single primary endpoint, we shall assume that an increase in the mean value of the response indicates a positive effect on the illness. Thus the new treatment giving a higher value than the control implies the new treatment

is better for patients. Conversely we could choose to look for a reduction in the observed endpoint meaning an improvement for the patient.

Initially we shall assume that the endpoint is observed immediately, or at least that any delay in observation is negligible. For example we might observe some measure of the illness before giving the treatment and then again 24 hours later, this allows us to assess the difference the treatment has made in a very short amount of time. The assumption of immediate response is made in Chapters 3, 4, 6 and 7. In Chapter 5 we remove this assumption working with delayed responses and discuss the application to survival endpoints.

Defining the average improvement provided by the new treatment as μ_x and the average improvement provided by the control as μ_y . It is μ_x and μ_y that must be compared by the trial, if the new treatment is better than the control treatment then we expect to see $\mu_x > \mu_y$. Proving $\mu_x > \mu_y$ shows that the new treatment is superior to that given in the control.

An alternative representation of the new treatment being superior to the control is to consider only the difference between the μ_x and μ_y , we will refer to this as the treatment effect denoted by θ , where $\theta = \mu_x - \mu_y$. Conducting the trial to assess whether $\theta > 0$ is equivalent to assessing whether $\mu_x > \mu_y$. We will formulate the questions of interest in terms of θ as this will allow for a clearer notation, particularly in the more complicated examples of the later chapters.

From the treatment effect we define the null hypothesis $H_0 : \theta \leq 0$, with the alternative hypothesis being $H_1 : \theta > 0$. The trial is then conducted to provide evidence to perform a hypothesis test with type I error rate $= \alpha$. Where we control the type I error at level α when,

$$\mathbb{P}_\theta(\text{Reject } H_0) \leq \alpha \quad \text{for all } \theta \leq 0. \quad (2.1)$$

The type I error rate may be thought of as the probability that a false claim of efficacy is made and is used to protect patients from treatments that provide no benefit.

2.2 Conducting a simple trial

In conducting a trial, patients are recruited from the population of interest. This population is the group of people with the disease being investigated who should benefit from the new treatment, there may be additional eligibility criteria based on other factors such as the stage of the disease. The patients are randomised into two treatment groups, one group of patients will be given the new treatment the other group of patients will be given the control. Within each group we observe the response to treatment allowing us to estimate the treatment effect. In practice recruitment and randomisation are non-trivial aspects of a trial design, however they are not the focus of the work to follow; we shall assume that recruitment and randomisation are conducted in such a way as to ensure that the patients on each treatment arm are representative of the overall population and randomised evenly.

We assume a total of n patients are recruited, where n is chosen before the trial begins. Of these patients n_x are randomised to the new treatment and n_y to the control treatment; the ratio n_x/n_y is chosen before randomisation to achieve the required sample size in each treatment arm. The observations collected from the patients receiving the new treatment are given by X_1, \dots, X_{n_x} , and the observations from the patients receiving the control are given by Y_1, \dots, Y_{n_y} . We shall assume that these observations are independent and identically distributed with $X_i \sim N(\mu_x, \sigma_x^2)$ for all $i = 1, \dots, n_x$ and $Y_i \sim N(\mu_y, \sigma_y^2)$ for all $i = 1, \dots, n_y$.

From the observations of the patients response to the treatment we find estimates $\bar{X} = \frac{1}{n_x} \sum_{i=1}^{n_x} X_i$ and $\bar{Y} = \frac{1}{n_y} \sum_{i=1}^{n_y} Y_i$ for μ_x and μ_y respectively. The estimate of the treatment effect, $\hat{\theta}$ say, is given by $\hat{\theta} = \bar{X} - \bar{Y}$ and under the assumptions that $n_x = n_y$ and the treatment and control have a common variance, σ^2 say, the distribution of the treatment effect is

$$\hat{\theta} \sim N\left(\theta, \frac{4\sigma^2}{n}\right). \quad (2.2)$$

It is not always the case that such simple assumptions may be made about the observations, however these assumptions will be sufficient for our investigations.

The methods that follow for Adaptive Enrichment do not require these assumptions, all that is required is the ability to provide a p-value for hypothesis testing.

From our estimate of the treatment effect we construct a hypothesis test for the null hypothesis. We find the Z-value from the estimate of the treatment effect $\hat{\theta}$, Z say, which is given by

$$Z = \frac{\sqrt{n}\hat{\theta}}{2\sigma},$$

which has the corresponding one sided P-value

$$P = 1 - \Phi(Z).$$

Under the null hypothesis $Z \sim N(0, 1)$ and $P \sim Unif(0, 1)$ and we reject the null hypothesis at level α when $Z \geq \phi^{-1}(1 - \alpha)$ and $P \leq \alpha$. This ensures $\mathbb{P}_\theta(\text{Reject true } H_0) \leq \alpha$ for all $\theta \leq 0$.

If H_0 is rejected (and therefore H_1 is accepted) then there is evidence to suggest that the new treatment is effective for treating the disease when compared to that used in the control. Recall here that the control of the type I error was important in the context of a confirmatory trial, as this is one of the requirements for approval of the treatment by the regulatory bodies. With this in mind we will also require control of type I error when conducting Adaptive Enrichment trials.

Example

Let us consider a small example to illustrate how a trial of this simplified form might run. Our aim is to demonstrate that $\mu_x > \mu_y$ or $\theta > 0$, giving the null hypothesis $H_0 : \theta \leq 0$ which we test at $\alpha = 0.025$.

In this example we will use $\sigma^2 = 22^2$, from which we will select a suitable sample size. To choose the sample size we consider the power, $1 - \beta$ say, of our trial at a given important value δ where this is given by

$$\mathbb{P}(\text{Reject } H_0 | \theta = \delta) = 1 - \beta.$$

Defining Φ to be the CDF of the standard normal distribution then to achieve power $1 - \beta$ at a given value of the true treatment effect $\theta = \delta$ we require the sample size given by

$$n = 4\sigma^2 \left(\frac{\Phi^{-1}(\alpha) + \Phi^{-1}(1 - \beta)}{\delta} \right)^2.$$

Setting the type I error to be $\alpha = 0.025$ for a one sided test of the null hypothesis and aiming for $1 - \beta = 0.9$ at $\tau = 10$, gives a required sample size of $n \approx 200$.

Choosing the true values to be $\mu_x = 15$ and $\mu_y = 0$ we simulate a single trial. By simulation we get values x_1, \dots, x_{100} for the new treatment with $x_i \sim N(15, 22^2)$ for $i = 1, \dots, 100$ and y_1, \dots, y_{100} for the control with $y_i \sim N(0, 22^2)$ for $i = 1, \dots, 100$. Looking at the results from one particular simulation, we observe $\bar{x} = 17.18$ and $\bar{y} = -1.62$, this gives $\hat{\theta} = 18.80$. From this value of $\hat{\theta}$ we obtain the p-value for the one sided test of the null hypothesis $p = 0.001$ and so under this particular realisation of the trial the null hypothesis $H_0 : \theta \leq 0$ is rejected.

Simulating a single realisation of a trial is interesting, however we will be placing higher emphasis on the overall properties of each trial designs. In this example the key characteristic is the power, we may wish to consider what happens to the power over a range of possible values of θ . A power curve such as the one shown in Figure 2-1 will help us to understand how the trial will behave as we change characteristics governing the design of a trial. Here we observe the power over a range of possible values of θ , where for a given true value $\theta = \delta$ we achieve power $= 1 - \beta$.

Although we can perform the computation of power directly it is worth discussing now how these results can be simulated as we will do this regularly in subsequent examples. For any particular value of θ we may repeat the above simulation in order to find the proportion of times that the null hypothesis is rejected. Table 2.1 shows some results from such a simulation, this simulation used the same parameters as the rest of the example. For the given values of μ_x and μ_y we simulate m trials, from this we obtain an estimate $1 - \hat{\beta}$ as the proportion of the time the null hypothesis was rejected, we also find the standard

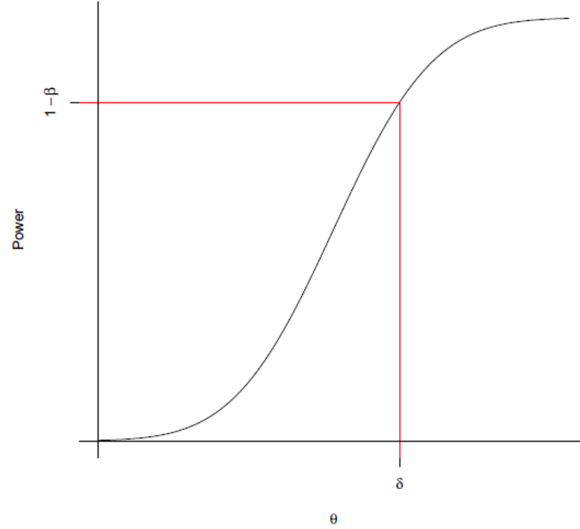


Figure 2-1: Example of a power curve

error for $1 - \hat{\beta}$ as $\sqrt{\frac{1}{n}(1 - \hat{\beta})\hat{\beta}}$. The table shows that each set of simulations have produced good estimates of power, as the number of simulations increase we see that the standard error of the estimate decreases. In the examples that follow we ensure we have run enough simulations to give our estimates the degree of accuracy we desire.

Assurance

In addition to understanding what happens over a range of possible values for θ we may also wish to consider the uncertainty about the true value of θ . Assurance allows us to do this; see O'Hagan et al. (2005). The uncertainty of the true parameter value is captured by defining a prior distribution for the true values, denote this by $\pi(\theta)$, the assurance is defined as the expected power over this prior distribution. Let $\nu(\theta)$ denote the power of the trial for a particular value of θ ,

$$\nu(\theta) = \Phi\left(\frac{\theta\sqrt{n}}{2\sigma} - \Phi^{-1}(\alpha)\right) \quad (2.3)$$

Assurance is given by,

$$\mathbb{E}_{\pi(\theta)}(\nu(\theta)) = \int_{-\infty}^{\infty} \pi(\theta)\nu(\theta)d\theta. \quad (2.4)$$

μ_x	μ_y	θ	m	$1 - \hat{\beta}$	Standard error	True power
10	10	0	1,000	0.021	0.005	0.025
10	10	0	10,000	0.025	0.002	0.025
10	10	0	100,000	0.025	0.001	0.025
10	5	5	1,000	0.355	0.015	0.362
10	5	5	10,000	0.371	0.005	0.362
10	5	5	100,000	0.364	0.002	0.362
10	0	10	1,000	0.901	0.010	0.895
10	0	10	10,000	0.892	0.003	0.895
10	0	10	100,000	0.896	0.001	0.895

Table 2.1: Demonstration of estimating power by simulating trials

Example

Returning to our previous example to add some uncertainty we choose the prior distribution for θ as $\theta \sim N(10, 4)$. From this prior distribution we evaluate Equation 2.4 to find $\mathbb{E}_{\pi(\theta)}(\nu(\theta)) = 0.86$.

If we had competing designs we could compare the value of assurance for each design and use this to inform our choice. We shall return to the idea of using Bayesian methods for comparing trial designs in Chapter 4.

Chapter 3

Conducting Adaptive Enrichment trials

3.1 Introducing sub-populations

3.1.1 Identification of sub-populations

The patient population to be used in the trial will have been carefully selected based on a set of inclusion criteria, despite this it is not always homogeneous. It is often possible to identify sub-populations that will react differently to the treatments, for our purposes we consider how this may impact on the treatment effect. It may be that part of the population has a particular characteristic that is expected to interact more favourably with the new treatment and hence this part of the population may receive a higher treatment effect. Thus we focus on trials that make use of this predictive information, such as those explained by the FDA draft guidance on Enrichment strategies of clinical trials, Food and Drug Administration and others (2012).

Our key assumption is that any sub-populations we use are identified before the trial begins and thus may be incorporated into the design of the trial. They may have been identified based on how the treatment is expected to work and characteristics that will assist this, or they may have been observed in previous stages of the development process or other trials for the disease.

Adaptive Enrichment trials aim to make efficient use of these pre-identified sub-populations by allowing for hypothesis testing of multiple sub-populations, selecting which should be used based on trial observations. When planning to test multiple hypotheses we must ensure we do not inflate the error rate, this requires the introduction of further methods in our hypothesis testing structure. We discuss these tools in simpler settings first before combining them. In Section 3.1.2 we formalise the questions of interest as the null hypotheses to be investigated, in Section 3.2 we discuss how the error rate can be controlled when multiple hypotheses are tested, in Section 3.4 we introduce combination tests as these are required when observing the data multiple times, in Section 3.5.1 we bring all of these tools together to see how a typical adaptive enrichment may be conducted.

3.1.2 Formulating the problem

We formulate our problem in a setting of exactly two sub-populations and only two hypotheses, however all of the methods described are easily extended to more hypotheses. We consider two scenarios for the sub-populations for conducting Adaptive Enrichment trials. The first is that there is a single sub-population of interest within a full population, where we expect the sub-population to react more favourably to the new treatment than the rest of the population. The second is that there are two distinct sub-populations that are thought to react favourably to the new treatment but it is not known which will receive the greater benefit.

It is possible to consider other scenarios for the sub-populations. For example if sub-populations are defined by the presence of some gene then a patient may belong to multiple sub-populations, giving a correlation between sub-populations. The two scenarios we consider allow us to investigate the performance of Adaptive Enrichment trials in a simple setting, while demonstrating how the methods may be applied across a range of formulations for the sub-populations.

Single sub-population

Under the first scenario it is believed that a single sub-population within the full population may receive a greater benefit from the new treatment. This is

the first formulation of our problem, we will refer to the sub-population as the first sub-population or the sub-population of interest. Our trials will attempt to answer two questions. Is the new treatment effective in the sub-population of interest? Is the new treatment effective in the population as a whole? To answer these questions a trial must be conducted that investigates both the treatment effect in the sub-population, θ_1 say, and the treatment effect in the full population, θ_3 say.

The treatment effect in each population may be found as it was in Section 2.2. In the sub-population of interest let μ_{x1} denote the average response in the patients who receive the new treatment and μ_{y1} denote the average response in the patients who receive the control treatment, then $\theta_1 = \mu_{x1} - \mu_{y1}$. Consider also the complement of the sub-population in this second sub-population let μ_{x2} denote the average response in the patients who receive the new treatment and μ_{y2} denote the average response in the patients who receive the control treatment. The treatment effect in the complement of the sub-population of interest, θ_2 say, is defined as $\theta_2 = \mu_{x2} - \mu_{y2}$.

To answer the question of whether the treatment is beneficial for all patients we define the treatment effect in the full population, θ_3 say. We combine the treatment effects from each sub-population to find the treatment effect in the full population. Let λ be the proportion of the full population that this sub-population of interest makes up, it follows that $\theta_3 = \lambda\theta_1 + (1 - \lambda)\theta_2$.

From the treatment effects we define two null hypotheses that we wish to test, $H_{01} : \theta_1 \leq 0$ in the sub-population of interest and $H_{03} : \theta_3 \leq 0$ in the full population; note that the full population may represent the same patients as the trial population we used in Section 2.2, in which case H_{03} is equivalent to the H_0 we used previously. It is possible to conduct a trial similarly to Section 2.2 that allows testing of both H_{01} and H_{03} or we could choose to conduct the trial only in the sub-population. Adaptive Enrichment provides a compromise between these two options in order to make efficient use of the sub-populations.

Double sub-population

Under the second scenario where there are two separate sub-populations, there are two key questions for any trial under this formulation of the problem. Is the new treatment effective in the first population? Is the new treatment effective in the second population? In this setting we do not test a null hypothesis for the full population. We refer to this as the second formulation of the problem. As before, a trial may be conducted to investigate the treatment effects in each population, θ_1 in the first population and θ_2 in the second population. Where sub-population 1 accounts for λ of the total sample and sub-population 2 accounts for $1 - \lambda$ of the total sample where $\lambda \in (0, 1)$.

As with the first formulation of the problem we may find estimates of the treatment effects in the two sub-populations, $\hat{\theta}_1$ and $\hat{\theta}_2$ in the first and second sub-populations respectively. From these treatment effects we define the null hypotheses $H_{01} : \theta_1 \leq 0$ in the first sub-population and $H_{02} : \theta_2 \leq 0$ in the second sub-population. As previously we may choose to conduct trials using either one of the sub-populations or using both, recruiting the trial from the appropriate sub-populations and testing the corresponding null hypotheses. The Adaptive Enrichment trial provides a compromise between these fixed sampling methods in order to make efficient use of the sub-populations.

3.2 Testing multiple hypotheses

3.2.1 Familywise error rate

Under both formulations of the problem we wish to test multiple null hypotheses. The testing of multiple hypotheses in the setting of a confirmatory clinical trial requires more careful definition of the error rate. The FamilyWise Error Rate (FWER) is defined as the probability of rejecting one or more true null hypotheses, in our case we require strong control of the FWER. Define $\boldsymbol{\theta}$ as a vector of the true treatment effects then we have strong control of the FWER at level α if

$$\mathbb{P}_{\boldsymbol{\theta}}(\text{Reject at least one true null hypothesis}) \leq \alpha \quad \text{for all } \boldsymbol{\theta}. \quad (3.1)$$

Dmitrienko et al. (2013) discuss techniques appropriate for strong control of the FWER, noting that the need to address multiplicity has been recognised by ICH Steering Committee and others (1998) and regulatory bodies Food and Drug Administration and others (1998) and Committee for Proprietary Medicinal Products and others (2002).

Example

To demonstrate FWER consider the hypotheses presented in section 3.1.2. Under the second formulation of the problem we test both $H_{01} : \theta_1 \leq 0$ and $H_{02} : \theta_2 \leq 0$. Given all possible values for $\boldsymbol{\theta} = (\theta_1, \theta_2)$ there are four possible combinations of true and false null hypotheses: both H_{01} and H_{02} are true, H_{01} is true and H_{02} is false, H_{01} is false and H_{02} is true, and finally both H_{01} and H_{02} are false. Note that changing 2 to 3 in the index of the second hypothesis yields identical results for the first formulation of the problem.

Under each combination of true and false null hypotheses we may state the conditions for strong control of the FWER. If both H_{01} and H_{02} are true then we require,

$$\mathbb{P}_{\boldsymbol{\theta}}(\text{Reject } H_{01}, H_{02} \text{ or both}) \leq \alpha.$$

If H_{01} is true and H_{02} is false then we require,

$$\mathbb{P}_{\boldsymbol{\theta}}(\text{Reject } H_{01}) \leq \alpha.$$

Simillarly if H_{01} is false and H_{02} is true then we require,

$$\mathbb{P}_{\boldsymbol{\theta}}(\text{Reject } H_{02}) \leq \alpha$$

If both H_{01} and H_{02} are false then no false claim can be made and so no condition is necessary. If all of these conditions are met by a particular method for hypothesis testing then we have strong control of the FWER as defined in equation 3.1.

3.2.2 Closed testing procedures

Gabriel (1969) discusses simultaneous test procedures for making multiple comparisons building on the work of Tukey (1951). Marcus et al. (1976) build on this in the form of a closed testing procedure. A closed testing procedure can be used to test a general number of null hypotheses while ensuring strong control of the FWER. Suppose we have $n \in \mathbb{N}$ treatment effects, $\theta_1, \dots, \theta_n$ say, which define the null hypotheses $H_{0i} : \theta_i \leq 0$ for $i = 1, \dots, n$. In addition to the individual hypotheses we define all the possible intersections of these null hypotheses from $H_{01} \cap H_{02}$ to $\bigcap_{i=1}^n H_{0i}$. The intersection between hypotheses, say $H_{0i} \cap H_{0j}$ for $i \neq j$, is given by

$$H_{0i} \cap H_{0j} = \theta_i \leq 0 \cap \theta_j \leq 0.$$

Suppose we are able to define level α tests for all possible intersection hypotheses then to reject any null hypothesis, say $H_{0i} : \theta_i \leq 0$ for $i = 1, \dots, n$, globally at level α we must reject all of the level α tests for hypotheses in which θ_i appears.

Example

Suppose we wish to conduct hypothesis test for the null hypotheses $H_{01} : \theta_1 \leq 0$ and $H_{02} : \theta_2 \leq 0$ whilst strongly controlling the FWER. A closed testing procedure for this pair of hypotheses will require level α tests of H_{01} , H_{02} and $H_{01} \cap H_{02}$. In order to reject H_{01} globally at level α the individual tests of H_{01} and $H_{01} \cap H_{02}$ must both be rejected. Similarly in order to reject H_{02} globally at level α the individual tests of H_{02} and $H_{01} \cap H_{02}$ must both be rejected.

Recall from equation 3.1 that in order to achieve strong control of the FWER we require,

$$\mathbb{P}_{\boldsymbol{\theta}}(\text{Reject at least one true null hypothesis}) \leq \alpha \text{ for all } \boldsymbol{\theta}.$$

Let $\boldsymbol{\theta} = (\theta_1, \dots, \theta_n)$, for any choice of $\boldsymbol{\theta}$ there is a corresponding set of true null hypotheses. We define \mathcal{H} as the vector of indexes corresponding to the true null

hypotheses, thus the intersection of all true null hypotheses is

$$\bigcap_{i \in \mathcal{H}} H_{0i}.$$

Under a closed testing procedure to reject any individual null hypothesis we must reject $\bigcap_{i \in \mathcal{H}} H_{0i}$ in order to reject any of the true null hypotheses globally. From the definition of the closed testing procedure we have

$$\mathbb{P}_{\boldsymbol{\theta}}(\text{Reject } \bigcap_{i \in \mathcal{H}} H_{0i}) \leq \alpha \quad \text{for all } \boldsymbol{\theta},$$

and hence we have strong control of the FWER.

Example

Returning to our two hypothesis example we can check that the conditions required for strong control of the FWER are met. If both H_{01} and H_{02} are true then the level α test of $H_{01} \cap H_{02}$ must be rejected to reject either or both of the true null hypotheses globally and since $H_{01} \cap H_{02}$ is true we have $\mathbb{P}_{\boldsymbol{\theta}}(\text{Reject } H_{01}) \leq \alpha$, $\mathbb{P}_{\boldsymbol{\theta}}(\text{Reject } H_{02}) \leq \alpha$ and $\mathbb{P}_{\boldsymbol{\theta}}(\text{Reject } H_{01} \text{ and } H_{02}) \leq \alpha$. If H_{01} is true and H_{02} is false then the level α test of H_{01} must be rejected to reject the true null hypothesis and we have $\mathbb{P}_{\boldsymbol{\theta}}(\text{Reject } H_{01}) \leq \alpha$. Similarly if H_{01} is false and H_{02} is true then the level α test of H_{02} must be rejected to reject the true null hypothesis and so $\mathbb{P}_{\boldsymbol{\theta}}(\text{Reject } H_{02}) \leq \alpha$. With all of these conditions met strong control of the FWER is achieved.

3.2.3 Procedures that give strong control of the FWER are closed testing procedures

Testing of multiple hypotheses is a problem that can be approached in many different ways, however I will only consider the use of closed testing procedures. This does not restrict the use of the trial designs presented as we claim that any testing procedure that gives strong control of the FWER for multiple hypotheses can be written as a closed testing procedure.

Proposition 3.2.1. *A hypothesis testing procedure gives strong control of the FWER if and only if it is a closed testing procedure.*

To see that this is true consider any testing procedure that gives strong control of FWER for $n \in \mathbb{N}$ null hypotheses, recalling equation 3.1 this means that

$$\mathbb{P}_{\boldsymbol{\theta}}(\text{Reject at least one true null hypothesis}) \leq \alpha \quad \text{for all } \boldsymbol{\theta}.$$

Let ξ_1, \dots, ξ_n be the rejection regions say corresponding to globally rejecting the null hypotheses H_{01}, \dots, H_{0n} . Let X be the observed data from the trial then we reject the null hypothesis $H_{0i} : \theta_i \leq 0$ globally at level α if $X \in \xi_i$. So from strong control of the FWER we have

$$\mathbb{P}_{\boldsymbol{\theta}}(\hat{\boldsymbol{\theta}} \in \xi_i) \leq \alpha$$

for all $\boldsymbol{\theta}$ such that H_{0i} is true. These rejection regions may be used to form the closed testing procedure for testing all the null hypotheses.

Strong control of the FWER ensures $\mathbb{P}_{\theta_i}(\text{Reject } H_{0i}) \leq \alpha$ for all $\boldsymbol{\theta}$ where $\theta_i \leq 0$, for $i = 1, \dots, n$ so the tests of individual hypotheses are at level α as required for a closed testing procedure.

For any subset \mathcal{I} of $1, \dots, n$ and associated intersection null hypothesis $\mathcal{H}_{\mathcal{I}} = \cap_{i \in \mathcal{I}} H_{0i}$ we require a level α test for the closed testing procedure. We define the rejection region for the test of $\mathcal{H}_{\mathcal{I}}$ as $\mathcal{R}_{\mathcal{I}} = \cup_{i \in \mathcal{I}} \xi_i$. For example for $\mathcal{I} = (1, 2)$ then the intersection null hypothesis is $\mathcal{H}_{\mathcal{I}} = H_{01} \cap H_{02}$ and the rejection region is defined as $\mathcal{R}_{\mathcal{I}} = \xi_1 \cup \xi_2$. We have that $\mathbb{P}_{\boldsymbol{\theta}}(\text{Reject } \mathcal{H}_{\mathcal{I}}) = \mathbb{P}_{\boldsymbol{\theta}}(X \in \mathcal{R}_{\mathcal{I}})$ and by strong control of the FWER

$$\mathbb{P}_{\boldsymbol{\theta}}(X \in \mathcal{R}_{\mathcal{I}}) \leq \alpha \quad \text{for all } \boldsymbol{\theta} \text{ for which } \mathcal{H}_{\mathcal{I}} \text{ is true}$$

since $X \in \mathcal{R}_{\mathcal{I}}$ implies a Familywise error is committed. So this rejection region provides a level α test for the intersection of the null hypothesis $\mathcal{H}_{\mathcal{I}}$.

By construction $X \in \xi_i$ implies that $X \in \mathcal{R}_{\mathcal{I}}$ for every $\mathcal{H}_{\mathcal{I}}$ with $i \in \mathcal{I}$. So using the tests of the individual hypotheses defined by the ξ_i and the intersection

hypotheses as we have defined above as a closed testing procedure is equivalent to the original multiple testing procedure that gave strong control of the FWER.

3.2.4 Consonance

Even in the simple case of two hypotheses a closed testing procedure requires rejection of multiple hypothesis tests to reject a single hypothesis globally. If a test of an intersection hypothesis can be rejected when the hypotheses involved may not be rejected globally the testing procedure is inefficient, this is the concept of consonance, see Romano et al. (2011). In the two hypothesis case we have rejection regions ξ_1 , ξ_2 and ξ_{12} for testing the null hypotheses H_{01} , H_{02} and $H_{01} \cap H_{02}$ the closed testing procedure is said to be consonant if

$$\xi_{12} \subseteq \xi_1 \cup \xi_2.$$

That is any test of intersection hypothesis should not be rejected where no hypotheses may be rejected globally, this property ensures that the testing procedure spends the full α efficiently. In our two hypothesis example if there are outcomes where $X \in \xi_{12}$ and $X \notin \xi_1 \cup \xi_2$ then the type I error rate for testing $H_{01} \cap H_{02}$ is not used efficiently.

3.3 Fixed sampling designs

3.3.1 Formulation 1: single sub-population

Before we describe the Adaptive Enrichment designs we consider some fixed sampling designs that will allow testing of the same null hypotheses. Recall from section 3.1.2 that under the first formulation of the problem the null hypotheses are $H_{01} : \theta_1 \leq 0$ and $H_{03} : \theta_3 \leq 0$. Where θ_1 is the treatment effect in the sub-population of interest and θ_3 is the treatment effect in the full population. The conduct of the trial is similar to that seen in section 2.2, however now we must keep track of which patients belong to each of the sub-populations.

Fixing the total sample size to be a total of n patients the recruitment is split between the sub-population of interest and its complement. We denote the

proportion of patients recruited from the sub-population of interest by λ and so the number of patients in the first sub-population is

$$n_1 = \lambda n.$$

Similarly the number of patients in the complement is

$$n_2 = (1 - \lambda)n$$

and $n = n_1 + n_2$. Within each sub-population patients are randomised between the new treatment and the control. In the first sub-population n_{x1} is the number of patients allocated to the new treatment and n_{y1} is the number of patients allocated to the control treatment. Likewise in the second sub-population n_{x2} is the number of patients allocated to the new treatment and n_{y2} is the number of patients allocated to the control treatment.

When collecting our observations we introduce an additional subscript to indicate the sub-population the patient is from. $X_{1,1}, \dots, X_{n_{x1},1}$ are the observations from patients in the sub-population of interest receiving the new treatment and $X_{1,2}, \dots, X_{n_{x2},2}$ are the observations from the patients in the complement population receiving the new treatment. Similarly $Y_{1,1}, \dots, Y_{n_{y1},1}$ are the observations from patients in the sub-population of interest receiving the control and $Y_{1,2}, \dots, Y_{n_{y2},2}$ are the observations from the patients in the complement population receiving the control.

From these observations estimates of μ_{x1} , μ_{x2} , μ_{y1} and μ_{y2} in the usual way, $\bar{X}_1 = \frac{1}{n_{x1}} \sum_{i=1}^{n_{x1}} X_{i,1}$, $\bar{X}_2 = \frac{1}{n_{x2}} \sum_{i=1}^{n_{x2}} X_{i,2}$, $\bar{Y}_1 = \frac{1}{n_{y1}} \sum_{i=1}^{n_{y1}} Y_{i,1}$ and $\bar{Y}_2 = \frac{1}{n_{y2}} \sum_{i=1}^{n_{y2}} Y_{i,2}$ respectively. These estimates can then be used to obtain estimates for θ_1 and θ_3 , we start by finding the treatment effect in each sub-population $\hat{\theta}_1 = \bar{X}_1 - \bar{Y}_1$, $\hat{\theta}_2 = \bar{X}_2 - \bar{Y}_2$ and the treatment effect in the full population is given by $\hat{\theta}_3 = \lambda \hat{\theta}_1 + (1 - \lambda) \hat{\theta}_2$. As we did in Section 2.2 we assume a common variance, σ^2 say, and split the sample evenly between the new treatment and the control. Under these assumptions the distributions for the treatment effects are

given by,

$$\hat{\theta}_1 \sim N\left(\theta_1, \frac{4\sigma^2}{\lambda n}\right), \hat{\theta}_2 \sim N\left(\theta_2, \frac{4\sigma^2}{(1-\lambda)n}\right), \hat{\theta}_3 \sim N\left(\theta_3, \frac{4\sigma^2}{n}\right) \quad (3.2)$$

The Z-values for testing the null hypotheses $H_{01} : \theta_1 \leq 0$ and $H_{03} : \theta_3 \leq 0$ are

$$Z_1 = \frac{\sqrt{\lambda n} \hat{\theta}_1}{2\sigma}, Z_3 = \frac{\sqrt{n} \hat{\theta}_3}{2\sigma}$$

respectively. The corresponding P-values are

$$P_1 = 1 - \Phi(Z_1), P_3 = 1 - \Phi(Z_3).$$

If $\theta_1 = 0$, $Z_1 \sim N(0, 1)$ and $P_1 \sim \text{Unif}(0, 1)$ and we reject $H_{01} : \theta_1 \leq 0$ when $Z_1 \geq \Phi^{-1}(1 - \alpha)$ which is equivalent to $P_1 \leq \alpha$. This ensures $\mathbb{P}_{\theta_1}(\text{Reject } H_{01}) \leq \alpha$ for all $\theta_1 \leq 0$. Similarly if $\theta_3 = 0$, $Z_3 \sim N(0, 1)$ and $P_3 \sim \text{Unif}(0, 1)$ and we reject $H_{03} : \theta_3 \leq 0$ when $Z_3 \geq \Phi^{-1}(1 - \alpha)$ which is equivalent to $P_3 \leq \alpha$. This ensures $\mathbb{P}_{\theta_3}(\text{Reject } H_{03}) \leq \alpha$ for all $\theta_3 \leq 0$. To apply the closed testing procedure we still require a test of the intersection of these null hypotheses.

3.3.2 Formulation 2: double sub-population

Under the second formulation of the problem presented in Section 3.1.2 we define the null hypotheses to be $H_{01} : \theta_1 \leq 0$ and $H_{02} : \theta_2 \leq 0$. Recruitment and randomisation within each sub-population proceeds in exactly the same way as seen in Section 3.3.1.

As before, with λ as the proportion of the sample in the first sub-population the sample in sub-population 1 is $n_1 = \lambda n$ and the sample in sub-population 2 is $n_2 = (1 - \lambda)n$. Assuming as before that exactly half of the patients are randomised to each treatment observations are collected in the same way, and with the same common variance the treatment effects follow the distributions given in Equation 3.2

$$\hat{\theta}_1 \sim N\left(\theta_1, \frac{4\sigma^2}{\lambda n}\right), \hat{\theta}_2 \sim N\left(\theta_2, \frac{4\sigma^2}{(1-\lambda)n}\right) \quad (3.3)$$

In this formulation of the problem we do not require an estimate of the treatment effect in the full population.

In the same way as before the Z-values for testing the null hypotheses $H_{01} : \theta_1 \leq 0$ and $H_{02} : \theta_3 \leq 0$ are

$$Z_1 = \frac{\sqrt{\lambda n} \hat{\theta}_1}{2\sigma}, \quad Z_2 = \frac{\sqrt{(1-\lambda)n} \hat{\theta}_2}{2\sigma}$$

respectively. The corresponding P-values are

$$P_1 = 1 - \Phi(Z_1), \quad P_2 = 1 - \Phi(Z_2).$$

If $\theta_1 = 0$, $Z_1 \sim N(0, 1)$ and $P_1 \sim \text{Unif}(0, 1)$ and we reject $H_{01} : \theta_1 \leq 0$ when $Z_1 \geq \Phi^{-1}(1-\alpha)$ which is equivalent to $P_1 \leq \alpha$. This ensures $\mathbb{P}_{\theta_1}(\text{Reject } H_{01}) \leq \alpha$ for all $\theta_1 \leq 0$. Similarly if $\theta_3 = 0$, $Z_3 \sim N(0, 1)$ and $P_3 \sim \text{Unif}(0, 1)$ and we reject $H_{02} : \theta_2 \leq 0$ when $Z_2 \geq \Phi^{-1}(1-\alpha)$ which is equivalent to $P_2 \leq \alpha$. This ensures $\mathbb{P}_{\theta_2}(\text{Reject } H_{02}) \leq \alpha$ for all $\theta_2 \leq 0$. Again to apply the closed testing procedure we still require a test of the intersection of these null hypotheses.

3.3.3 Simes method for testing the intersection hypothesis

In the closed testing procedure we require a test of the intersection hypothesis. The method introduced by Simes (1986) allows the construction of a level α test for an intersection hypothesis with less conservatism than the Bonferroni correction. Suppose we wish to test $H_{01} \cap H_{02}$, let P_1 and P_2 be the p-values for testing H_{01} and H_{02} respectively, Simes method defines the p-value for testing $H_{01} \cap H_{02}$ to be

$$P_{12} = \min(2\min(P_1, P_2), \max(P_1, P_2)). \quad (3.4)$$

When $\theta_1 = 0$ and $\theta_2 = 0$ P_1 and P_2 are independent with P_1 and $P_2 \sim \text{Unif}(0, 1)$. To reject $H_1 \cap H_2$ we must either have $2\min(P_1, P_2) \leq \alpha$ which requires P_1 or $P_2 \leq \alpha/2$ or $\max(P_1, P_2) \leq \alpha$ which requires P_1 and $P_2 \leq \alpha$.

Taking the minimum of these means that $\mathbb{P}(P_{12} \leq \alpha)$ may be written as,

$$\mathbb{P}(P_{12} \leq \alpha) = \mathbb{P}((P_1 \cap P_2 \leq \alpha) \cup (P_1 \leq \alpha/2 \cap P_2 > \alpha) \cup (P_1 > \alpha \cap P_2 \leq \alpha/2)).$$

Using the independent uniform distributions under the null hypotheses we can see that,

$$\begin{aligned} & \mathbb{P}((P_1 \cap P_2 \leq \alpha) \cup (P_1 \leq \alpha/2 \cap P_2 > \alpha) \cup (P_1 > \alpha \cap P_2 \leq \alpha/2)) \\ &= \mathbb{P}(P_1 \cap P_2 \leq \alpha) + \mathbb{P}(P_1 \leq \alpha/2 \cap P_2 > \alpha) + \mathbb{P}(P_1 > \alpha \cap P_2 \leq \alpha/2) \\ &= \mathbb{P}(P_1 \leq \alpha)\mathbb{P}(P_2 \leq \alpha) + \mathbb{P}(P_1 \leq \alpha/2)\mathbb{P}(P_2 > \alpha) + \mathbb{P}(P_1 > \alpha)\mathbb{P}(P_2 \leq \alpha/2) \\ &= \alpha^2 + 2(\alpha/2)(1 - \alpha) \\ &= \alpha^2 + \alpha - \alpha^2 = \alpha \end{aligned}$$

and thus the probability we reject the null hypothesis $H_1 \cap H_2$ when $\theta_1 = 0$ and $\theta_2 = 0$ is α . In this case the argument is easily demonstrated graphically. The solid line in Figure 3-1 shows the boundary where the intersection p-value is at level α , $P_{12} = \alpha$, below this line $P_{12} < \alpha$, the dashed lines show the corresponding boundaries where $P_1 = \alpha$ and $P_2 = \alpha$. The areas for rejecting the null hypotheses $H_{01} \leq 0$ and $H_{02} \leq 0$ are α , assuming uniform probability in this square it is clear to see the total area for rejecting the intersection $H_{01} \cap H_{02}$ is also α . Consonance is also easily observed as whenever the intersection is rejected at least one of the individual hypotheses is also rejected.

When $\theta_1 < 0$ and $\theta_2 < 0$ the proof requires more detail; the p_{ctud} condition introduced by Brannath et al. (2002) in the context of combination tests is that the P-values are stochastically larger than or equal to the $\text{Unif}(0, 1)$ and could be used for this proof, we omit the full detail here. Likewise the proof that Simes method is appropriate under the first formulation of our problem is more detailed as we do not have independence between the estimates $\hat{\theta}_1$ and $\hat{\theta}_3$. A general proof of the suitability of Simes method is given by Sarkar and Chang (1997) and Sarkar (1998), this proof applies in this case as it requires a positive association between the variables which is given by the positive correlation.

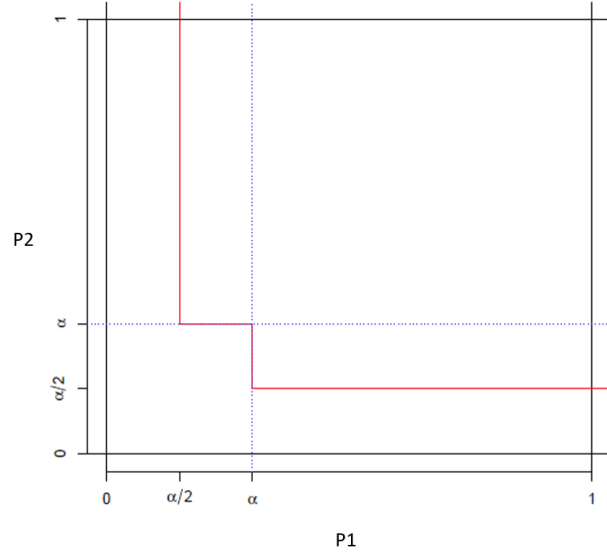


Figure 3-1: Demonstrating Simes method

Simes method is consonant. If the intersection hypothesis $H_{01} \cap H_{02}$ is rejected at level α then we show that either H_{01} or H_{02} must have been rejected. If we reject $H_{01} \cap H_{02}$ we have

$$P_{12} \leq \alpha$$

this gives

$$\min(2\min(P_1, P_2), \max(P_1, P_2)) \leq \alpha,$$

so either

$$2\min(P_1, P_2) \leq \alpha$$

or

$$\max(P_1, P_2) \leq \alpha.$$

If the intersection is rejected due to $2\min(P_1, P_2) \leq \alpha$ then $P_i \leq \alpha/2$ for $i = 1, 2$ and so either H_{01} or H_{02} must have been rejected. If the intersection is rejected

due to $\max(P_1, P_2) \leq \alpha$ then both $P_1 \leq \alpha$ and $P_2 \leq \alpha$ and so both H_{01} and H_{02} must have been rejected. This argument also applies under the first formulation of our problem if we change the subscript 2 to 3.

3.3.4 Example: Formulation 2

A single trial

We now extend the example from Chapter 2 into two sub-populations. Under the second formulation of the problem we test the null hypotheses $H_{01} : \theta_1 \leq 0$ and $H_{02} : \theta_2 \leq 0$ while strongly controlling the FWER at $\alpha = 0.025$. Choosing $\lambda = 1/2$ we use the same variance and total sample size from Section 2.2, $\sigma^2 = 22^2$ and $n = 200$. Splitting this sample between each sub-population and treatment group gives $n_{x1} = n_{x2} = n_{y1} = n_{y2} = 50$. We could change the proportion of the sample in each sub-population and hence the sample sizes based on desired power for each hypothesis as we did before, however in order to make comparisons between different trial designs we will keep this constant.

For this example we will also assume both populations receive the same benefit from the new treatment. Setting the true treatment effects $\mu_{x1} = \mu_{x2} = 10$ for the new treatment and $\mu_{y1} = \mu_{y2} = 0$ we may simulate an example trial. By simulation we get observations: $x_{i,1} \sim N(10, 22^2)$ for $i = 1, \dots, 50$ in the first sub-population for the new treatment, $x_{i,2} \sim N(10, 22^2)$ for $i = 1, \dots, 50$ in the second sub-population for the new treatment, $y_{i,1} \sim N(0, 22^2)$ for $i = 1, \dots, 50$ in the first sub-population for the control and $y_{i,2} \sim N(0, 22^2)$ for $i = 1, \dots, 50$ in the second sub-population for the control.

Suppose for one particular realisation in population 1 we observe $\bar{x}_1 = 13.5$ and $\bar{y}_1 = -3.7$ giving $\hat{\theta}_1 = 17.2$, and in population 2 we observe $\bar{x}_2 = 5.3$ and $\bar{y}_2 = 0.1$ giving $\hat{\theta}_2 = 5.2$. This gives p-values $p_1 < 0.001$ and $p_2 = 0.119$ for testing the null hypotheses in sub-population 1 and 2 respectively, furthermore using Simes method we find the intersection p-value $p_{12} < 0.001$. In this realisation of the trial we are able to reject H_{01} since both $p_1 < \alpha$ and $p_{12} < \alpha$ however we accept H_{02} since $p_2 > \alpha$.

Overall characteristics

When comparing trial designs we want to understand the operating characteristics of each trial. Given the two null hypotheses we measure the performance of the trial by the probabilities of rejecting each hypothesis individually or both hypotheses at the same time, these are $\mathbb{P}(\text{reject } H_{01} \text{ and accept } H_{02})$, $\mathbb{P}(\text{accept } H_{01} \text{ and reject } H_{02})$ and $\mathbb{P}(\text{reject both } H_{01} \text{ and } H_{02})$. To reject H_{01} we must observe $p_1 \leq \alpha$ and $p_{12} \leq \alpha$, similarly to reject H_{02} we must observe $p_2 \leq \alpha$ and $p_{12} \leq \alpha$ and to reject both H_{01} and H_{02} we must observe $p_1 \leq \alpha$, $p_2 \leq \alpha$ and $p_{12} \leq \alpha$.

Noting that under Simes rule if $p_2 > \alpha$ then for $p_{12} \leq \alpha$ we must have $p_1 \leq \alpha/2$ and vice versa. Thus if $p_1 = \alpha$ when $\hat{\theta}_1 = k_{1,\alpha}$ and $p_2 = \alpha$ when $\hat{\theta}_2 = k_{2,\alpha}$ we may compute the probabilities of rejecting the null hypotheses as follows:

$$\begin{aligned} \mathbb{P}(\text{reject } H_{01} \text{ and accept } H_{02}) &= \mathbb{P}(p_1 \leq \alpha \cap p_2 > \alpha \cap p_{12} \leq \alpha) \\ &= \mathbb{P}(p_1 \leq \alpha/2 \cap p_2 > \alpha) \\ &= \mathbb{P}(\hat{\theta}_1 \geq k_{1,\alpha/2} \cap \hat{\theta}_2 < k_{2,\alpha}), \end{aligned}$$

$$\begin{aligned} \mathbb{P}(\text{accept } H_{01} \text{ and reject } H_{02}) &= \mathbb{P}(p_1 > \alpha \cap p_2 \leq \alpha \cap p_{12} \leq \alpha) \\ &= \mathbb{P}(p_1 > \alpha \cap p_2 \leq \alpha/2) \\ &= \mathbb{P}(\hat{\theta}_1 < k_{1,\alpha} \cap \hat{\theta}_2 \geq k_{2,\alpha/2}), \end{aligned}$$

$$\begin{aligned} \mathbb{P}(\text{reject both } H_{01} \text{ and } H_{02}) &= \mathbb{P}(p_1 \leq \alpha \cap p_2 \leq \alpha \cap p_{12} \leq \alpha) \\ &= \mathbb{P}(p_1 \leq \alpha \cap p_2 \leq \alpha) \\ &= \mathbb{P}(\hat{\theta}_1 \geq k_{1,\alpha} \cap \hat{\theta}_2 \geq k_{2,\alpha}). \end{aligned}$$

Furthermore in this example we have independence of $\hat{\theta}_1$ and $\hat{\theta}_2$ so

$$\begin{aligned}\mathbb{P}(\text{reject } H_{01} \text{ and accept } H_{02}) &= \mathbb{P}(p_1 \leq \alpha \cap p_2 > \alpha \cap p_{12} \leq \alpha) \\ &= \mathbb{P}(\hat{\theta}_1 \geq k_{1,\alpha/2})\mathbb{P}(\hat{\theta}_2 < k_{2,\alpha})\end{aligned}$$

$$\begin{aligned}\mathbb{P}(\text{accept } H_{01} \text{ and reject } H_{02}) &= \mathbb{P}(p_1 > \alpha \cap p_2 \leq \alpha \cap p_{12} \leq \alpha) \\ &= \mathbb{P}(\hat{\theta}_1 < k_{1,\alpha})\mathbb{P}(\hat{\theta}_2 \geq k_{2,\alpha/2})\end{aligned}$$

and

$$\mathbb{P}(\text{reject } H_{01} \text{ and reject } H_{02}) = \mathbb{P}(\hat{\theta}_1 \geq k_{1,\alpha})\mathbb{P}(\hat{\theta}_2 \geq k_{2,\alpha})$$

which may all be evaluated directly when the distributions are known. From Equation 3.3 we have

$$\hat{\theta}_1 \sim N\left(\theta_1, \frac{4\sigma^2}{\lambda n}\right) \text{ and } \hat{\theta}_2 \sim N\left(\theta_2, \frac{4\sigma^2}{(1-\lambda)n}\right)$$

Table 3.1 shows how these probabilities vary as we vary the values of θ_1 and θ_2 . We see that the probability of falsely rejecting one or more true null hypotheses is never more than the nominal $\alpha = 0.025$, as expected. We also observe that the probability of rejecting each null hypothesis depends not only on the effect in that null hypothesis but also the effect in the other, this is due to the additional rigour required under multiple testing. Notice that even when both treatment effects are 10 (which is what the original sample size was based on) the probability of rejecting at least one null hypothesis is reduced to 0.774 from the 0.895 seen in example from section 2.2 where a single hypothesis is tested.

3.3.5 Example: Formulation 1

A single trial

Under the first formulation of our problem we consider a sub-population within a full population. The key difference with the previous example is that our hypothesis testing is now concerned with the null hypotheses $H_{01} : \theta_1 \leq 0$ and $H_{03} : \theta_3 \leq 0$. Using $\lambda = 1/2$, $\sigma^2 = 22^2$ and $n = 200$ as we did previously we

θ_1	θ_2	$\mathbb{P}(\text{reject both})$	$\mathbb{P}(\text{reject } H_{01} \text{ only})$	$\mathbb{P}(\text{reject } H_{02} \text{ only})$	Σ
0	0	0.001	0.012	0.012	0.025
10	0	0.016	0.500	0.005	0.521
0	10	0.005	0.016	0.500	0.521
10	10	0.388	0.193	0.193	0.774
10	5	0.128	0.407	0.051	0.586
10	7.5	0.249	0.308	0.111	0.668

Table 3.1: Probabilities of rejecting null hypotheses in independent sub-populations using Simes' method for the intersection hypothesis

simulate the observations from a single trial in the same way as we did for the second formulation of the problem.

Using the same realisations for the observations as we did for formulation 2 we may test H_{01} and H_{03} . Recall that we observed $\bar{x}_1 = 13.5$ and $\bar{y}_1 = -3.7$ giving $\hat{\theta}_1 = 17.2$ in population 1, and we observed $\bar{x}_2 = 5.3$ and $\bar{y}_2 = 0.1$ giving $\hat{\theta}_2 = 5.2$ in population 2, combining the information from both sub-populations gives $\hat{\theta}_3 = 11.2$. This gives p-values $p_1 < 0.001$ as before and $p_3 < 0.001$ and using Simes method we find the intersection p-value $p_{13} < 0.001$. So we may reject both H_{01} and H_{03} since $p_1 < \alpha$, $p_3 < \alpha$ and $p_{13} < \alpha$.

Overall characteristics

To compute the operating characteristics of this trial we make use of the same computations as the previous example

$$\mathbb{P}(\text{reject } H_{01} \text{ and accept } H_{03}) = \mathbb{P}(\hat{\theta}_1 \geq k_{1,\alpha/2} \cap \hat{\theta}_2 < k_{3,\alpha}),$$

$$\mathbb{P}(\text{accept } H_{01} \text{ and reject } H_{03}) = \mathbb{P}(\hat{\theta}_1 < k_{1,\alpha} \cap \hat{\theta}_2 \geq k_{3,\alpha/2}),$$

$$\mathbb{P}(\text{reject both } H_{01} \text{ and } H_{03}) = \mathbb{P}(\hat{\theta}_1 \geq k_{1,\alpha} \cap \hat{\theta}_2 \geq k_{3,\alpha}).$$

θ_1	θ_2	θ_3	$\mathbb{P}(\text{reject both})$	$\mathbb{P}(\text{reject } H_{01} \text{ only})$	$\mathbb{P}(\text{reject } H_{03} \text{ only})$
0	0	0	0.008	0.007	0.007
10	-10	0	0.025	0.488	0.000
10	10	10	0.611	0.006	0.239
10	0	5	0.328	0.212	0.016
10	5	7.5	0.531	0.057	0.095

Table 3.2: Probabilities of rejecting null hypotheses in sub and full populations using Simes' method for the intersection hypothesis

Under this formulation of the problem we do not have independence of the treatment effects, however we do know that $\hat{\theta}_1$ and $\hat{\theta}_3$ follow the joint distribution,

$$\begin{pmatrix} \hat{\theta}_1 \\ \hat{\theta}_3 \end{pmatrix} \sim N_2 \left(\begin{pmatrix} \theta_1 \\ \theta_3 \end{pmatrix}, \begin{pmatrix} 4\sigma^2/\lambda n & 4\sigma^2/n \\ 4\sigma^2/n & 4\sigma^2/n \end{pmatrix} \right)$$

from which we may compute these probabilities (the covariance is given by $\text{cov}(\theta_1, \lambda\theta_1 + (1-\lambda)\theta_2) = \lambda\text{var}(\theta_1)$), we use the R package mvtnorm Genz et al. (2008).

Table 3.2 shows the operating characteristics of the trial under this formulation of the problem. As expected we see that the FWER is strongly controlled although there is some conservatism due to the correlation between the sub-populations. As before the probabilities of rejecting null hypotheses depend on both treatment effects. Under this formulation of the problem we see that the trials most often reject both null hypotheses, even when this effect is mostly driven by the sub-population of interest.

3.3.6 Fixed enrichment

Under either formulation of the problem we may conduct the trial entirely within a single sub-population; if we do this we are conducting a fixed Enrichment design as discussed under the predictive Enrichment designs discussed in Food and Drug Administration and others (2012). The fixed Enrichment design is conducted in exactly the same way as the trial described in section 2.2 with the trials patient population restricted to the selected sub-population.

Recall that the null hypothesis to be tested is of the form $H_0 : \theta \leq 0$, this could be $H_{01} : \theta_1 \leq 0$ under both formulations of the problem or $H_{02} : \theta_2 \leq 0$ under the second formulation of the problem. The same total of n patients will be recruited and randomised equally between the new treatment and the control.

As in Section 2.2 we collect observations from the treatment and control to estimate the treatment effect. Suppose we are conducting the trial in sub-population 1 then we find the estimate for the treatment effect follows the distribution

$$\hat{\theta}_1 \sim N\left(\theta_1, \frac{4\sigma^2}{n}\right).$$

As usual we find the Z-value from the estimate of the treatment

$$Z_1 = \frac{\sqrt{n}\hat{\theta}_1}{2\sigma},$$

which has the corresponding P-value

$$P_1 = 1 - \Phi(Z_1).$$

If $\theta_1 = 0$, $Z_1 \sim N(0, 1)$ and $P_1 \sim \text{Unif}(0, 1)$ and we reject $H_{01} : \theta_1 \leq 0$ when $Z_1 \geq \Phi^{-1}(1 - \alpha)$ which is equivalent to $P_1 \leq \alpha$. This ensures $\mathbb{P}_{\theta_1}(\text{Reject } H_{01}) \leq \alpha$ for all $\theta_1 \leq 0$.

Example

Suppose we restrict recruitment to sub-population i and test the corresponding null hypothesis $H_{0i} : \theta_i \leq 0$ for $i = 1, 2$. The probability of rejecting H_{0i} depends on θ_i in exactly the same way as the example of Section 2.2. Table 3.3 is a recap of the operating characteristics of this trial given true treatment effects, as expected we see an increase in power as the treatment effect rises.

3.4 Combination tests

Adaptive Enrichment designs involve an interim analysis where we may adapt the recruitment strategy for the remainder of the trial. Which data are collected after the interim analysis depends on the interim decision, and therefore the

μ_x	μ_y	θ_i	$\mathbb{P}_{\theta_i}(\text{Reject } H_{0i})$
10	10	0	0.025
10	5	5	0.362
10	2.5	7.5	0.674
10	0	10	0.895

Table 3.3: Power in a single hypothesis trial

data observed by this interim point. These multiple looks at the data must be accounted for when hypothesis testing in order to maintain strong control of the FWER.

Consider a single null hypothesis $H_0 : \theta \leq 0$ which is to be tested in a two stage experimental design. Suppose we split our sample such that we have $n^{(1)}$ observations from the first stage of the trial and $n^{(2)}$ observations from the second stage and we have a common variance σ^2 . We obtain estimates $\hat{\theta}^{(1)}$ and $\hat{\theta}^{(2)}$, for the first and second stage of the trial respectively. These estimates correspond to Z-values

$$Z^{(1)} = \frac{\sqrt{n^{(1)}}\hat{\theta}^{(1)}}{2\sigma}, Z^{(2)} = \frac{\sqrt{n^{(2)}}\hat{\theta}^{(2)}}{2\sigma}$$

respectively, for which the corresponding P-values are

$$P^{(1)} = 1 - \Phi(Z^{(1)}), P^{(2)} = 1 - \Phi(Z^{(2)}).$$

From this a combination test will find the corresponding combined Z-value, $Z^{(c)}$ say, and P-value, $P^{(c)}$ say. We require control of the type I error rate, recall from Equation 2.1 that this is the case when

$$\mathbb{P}_{\theta}(\text{Reject } H_0) \leq \alpha \quad \text{for all } \theta \leq 0.$$

To achieve this combination tests require that $Z^{(1)}$ and $Z^{(2)}$ are independent random variables, and thus $P^{(1)}$ and $P^{(2)}$ are also independent. When $\theta = 0$ this means $Z^{(i)} \sim N(0, 1)$ and $P^{(i)} \sim \text{Unif}(0, 1)$ for $i = 1, 2$ which ensures $\mathbb{P}_{\theta}(\text{Reject } H_0) \leq \alpha$ for $\theta = 0$. For our designs the data from the first stage $D^{(1)}$ may change how the second stage is conducted and so $Z^{(2)}|D^{(1)} = d^{(1)} \sim N(0, 1)$, however this is true for all realisations of the first stage $d^{(1)}$ giving us conditional independence between $Z^{(1)}$ and $Z^{(2)}$ and we have $Z^{(2)} \sim N(0, 1)$. For the case

when $\theta < 0$ Brannath et al. (2002) use the p_{clud} condition, requiring only that the P-values to be stochastically larger than or equal to the $\text{Unif}(0, 1)$.

3.4.1 Weighted inverse normal

For the time being we shall use the weighted inverse normal combination test, see Lehmacher and Wassmer (1999), this uses $Z^{(1)}$ and $Z^{(2)}$ to find $P^{(c)}$. The statistics $Z^{(1)}$ and $Z^{(2)}$ are weighted together to find the combined Z-value $Z^{(c)}$. With pre-specified weights w_1 and w_2 such that $w_1^2 + w_2^2 = 1$, we find

$$Z^{(c)} = w_1 Z^{(1)} + w_2 Z^{(2)}. \quad (3.5)$$

The combined Z-value corresponds to the combined p-value in the usual way

$$P^{(c)} = 1 - \Phi(Z^{(c)}).$$

When $\theta = 0$, $Z^{(c)}$ is a sum of two independent normal random variables and is therefore normally distributed. The choice of the weights give us that $Z^{(c)} \sim N(0, 1)$ and so $P^{(c)} \sim \text{Unif}(0, 1)$ which gives a level α test for H_0 . When $\theta < 0$ the P_{clud} argument applies, as in Brannath et al. (2002). Using either $Z^{(c)}$ or $P^{(c)}$ we construct a level α test for H_0 in the usual way.

The pre-specified weights w_1 and w_2 influence the probability of rejecting the null hypothesis of this procedure, if we achieve a sample such that the weights are in the same proportion as the sample size in each stage then the weighted inverse normal is as powerful a test for $H_0 : \theta \leq 0$ as testing without multiple stages since we achieve the same variance. Having pre-specified the weights achieving our sample such that

$$\frac{n^{(1)}}{n^{(2)}} = \frac{w_1}{w_2},$$

will be optimal, the consequence of this is that we choose our weights based on the expected sample size within each stage.

Example

We return again to the example from section 2.2, but now we split the trial into two stages. The pre-interim recruitment cohort will recruit $n^{(1)}$ patients and the post-interim recruitment cohort will recruit the remaining $n^{(2)}$ patients such that $n^{(1)} + n^{(2)} = n$. Splitting the trial such that we recruit half of the patients in the first stage of the trial we will first observe 50 patients from each treatment arm in stage 1 and then 50 patients from each treatment arm in stage 2.

We assume the observations are still from the same distributions but now we also index them by stage. So in the first stage we collect observations $x_i^{(1)} \sim N(10, 22^2)$, $i = 1, \dots, 50$, for the new treatment and $y_i^{(1)} \sim N(0, 22^2)$, $i = 1, \dots, 50$, for the control. Similarly in the second stage we collect observations $x_i^{(2)} \sim N(10, 22^2)$, $i = 1, \dots, 50$, for the new treatment and $y_i^{(2)} \sim N(0, 22^2)$, $i = 1, \dots, 50$, for the control. We obtain estimates of the treatment effect within each stage: $\hat{\theta}^{(1)}$ in the first stage and $\hat{\theta}^{(2)}$ in the second stage. These estimates are combined using the weighted inverse normal combination test to give an overall p-value for the null hypothesis.

For one particular simulation we observe $\hat{\mu}_x^{(1)} = 10.9$, $\hat{\mu}_y^{(1)} = 4.3$, $\hat{\mu}_x^{(2)} = 14.0$ and $\hat{\mu}_y^{(2)} = -3.8$. These give $\hat{\theta}^{(1)} = 6.6$ and $\hat{\theta}^{(2)} = 17.8$, to combine these we find the corresponding Z-values $Z^{(1)} = 1.5$ and $Z^{(2)} = 4.0$. Choosing weights $w_1 = w_2 = 1/\sqrt{2}$, applying these gives the combined Z-value $Z^{(c)} = 3.9$ which may be used to find the p-value $p^{(c)} < 0.001$ which allows us to accept/reject the null hypothesis for the two stage trial.

In Adaptive Enrichment trials we still have to pre-specify the weights for the overall testing procedure, however the proportions of the sample in the first and second stages will vary based on the choice of recruitment made at the interim analysis. Fixing an overall sample size Table 3.4 compares the power of the trial as this proportion varies when the weights for the weighted inverse normal are $1/\sqrt{2}$ and the treatment effect is 10. We see when the sample matches the weights at 0.50 the probability of rejecting the null hypothesis is 0.90 as planned for the single stage trial, this power drops off as the observed trial proportions move further from optimal. For our investigations of Adaptive Enrichment trials

Proportion in pre-interim recruitment cohort	$\mathbb{P}(\text{Reject } H_0)$
0.10	0.82
0.25	0.87
0.50	0.90
0.75	0.87
0.90	0.82

Table 3.4: Power when varying the first stage proportion

we will usually use these weights, with the proportion of the sample in the first stage being either 1/2 or 1/3. The results from Table 3.4 show that this choice of weights will perform reasonably well in both cases.

3.5 Conducting adaptive enrichment trials

3.5.1 Adaptive enrichment methodology

We now have all of the basic components required to conduct Adaptive Enrichment trials, all that remains is to put them together. Under either formulation of the problem the Adaptive Enrichment trial will begin in the same way as the fixed sampling method testing multiple hypotheses. Patients are initially recruited from all available populations, with the ability to be able to test both hypotheses as appropriate to the formulation being used. After a pre-determined amount of the total sample has been recruited, an interim analysis is conducted to decide which sub-populations will be sampled for the rest of the trial. The options at the interim analysis are given by the sub-populations as defined by the formulation: the trial may continue as it started (recruiting from all populations) or the trial may be enriched, that is the remainder of the observations are all recruited from one sub population. Wang and Hung (2013) give an overview of Adaptive Enrichment designs, with a case study of a confirmatory trial. Note that the hypothesis testing methods for the Adaptive Enrichment trial must be pre-defined, thus the null hypotheses are defined and the formulation of the problem is chosen before the trial begins.

Data collection

Taking the same total sample size as the fixed sample designs, n patients will be recruited overall, choosing $n^{(1)}$ as the number of patients to be recruited in the pre-interim cohort and $n^{(2)}$ as the number of patients to be recruited in the post-interim cohort. Whichever formulation is being used there are two sub-populations from which to recruit. The patients in the pre-interim cohort are split between the two sub-populations. If λ is the proportion of the total population represented by sub-population 1 then $n_1^{(1)} = \lambda n^{(1)}$ is the number of patients in the first sub-population in the pre-interim cohort and $n_2^{(1)} = (1 - \lambda)n^{(1)}$ is the number of patients in the second sub-population in the pre-interim cohort.

In the post interim cohort, we denote the number of patients from the first sub-population by $n_1^{(2)}$ and the number of patients from the second sub-population by $n_2^{(2)}$. If the decision at the interim analysis was to continue with all populations then the patients are split between the two sub-populations as they were in the first stage of the trial, so $n_1^{(2)} = \lambda n^{(2)}$ and $n_2^{(2)} = (1 - \lambda)n^{(2)}$. Under both formulations the trial may be enriched into the first sub-population, in which case $n_1^{(2)} = n^{(2)}$ and $n_2^{(2)} = 0$. For formulation 2 it is also possible to enrich into the second sub-population, in which case $n_1^{(2)} = 0$ and $n_2^{(2)} = n^{(2)}$.

Within each of these recruitment groups the patients are randomised equally between the new treatment and the control treatment as we did before. In the pre-interim recruitment cohort we collect observations as before: in the first subpopulation we observe $X_{i,1}^{(1)}$, $i = 1, \dots, n_1^{(1)}/2$, from those receiving the new treatment and $Y_{i,1}^{(1)}$, $i = 1, \dots, n_1^{(1)}/2$, from those receiving the control; in the second subpopulation we observe $X_{i,2}^{(1)}$, $i = 1, \dots, n_2^{(1)}/2$, from those receiving the new treatment and $Y_{i,2}^{(1)}$, $i = 1, \dots, n_2^{(1)}/2$, from those receiving the control.

From these observations estimates of the treatment effects are found as previously, giving us estimates of the treatment effect from the pre-interim cohort: $\hat{\theta}_1^{(1)}$ from sub-population 1, $\hat{\theta}_2^{(1)}$ from sub-population 2 and under the first formulation of the problem $\hat{\theta}_3^{(1)} = \lambda \hat{\theta}_1^{(1)} + (1 - \lambda) \hat{\theta}_2^{(1)}$. As before we use a

common variance, σ^2 , to write the distributions of these summary statistics.

$$\hat{\theta}_1^{(1)} \sim N\left(\theta_1, \frac{4\sigma^2}{\lambda n^{(1)}}\right),$$

$$\hat{\theta}_2^{(1)} \sim N\left(\theta_2, \frac{4\sigma^2}{(1-\lambda)n^{(1)}}\right)$$

and

$$\hat{\theta}_3^{(1)} \sim N\left(\theta_3, \frac{4\sigma^2}{n^{(1)}}\right).$$

These have corresponding Z-values

$$Z_1^{(1)} = \frac{\sqrt{\lambda n^{(1)}} \hat{\theta}_1^{(1)}}{2\sigma},$$

$$Z_2^{(1)} = \frac{\sqrt{(1-\lambda)n^{(1)}} \hat{\theta}_2^{(1)}}{2\sigma},$$

and

$$Z_3^{(1)} = \frac{\sqrt{n^{(1)}} \hat{\theta}_3^{(1)}}{2\sigma},$$

and P-values

$$P_1^{(1)} = 1 - \Phi\left(Z_1^{(1)}\right),$$

$$P_2^{(1)} = 1 - \Phi\left(Z_2^{(1)}\right),$$

and

$$P_3^{(1)} = 1 - \Phi\left(Z_3^{(1)}\right).$$

The recruitment of the post-interim cohort depends on the decision made at the interim analysis. If the trial continues in both sub-populations observations are collected in the same way as described above and we find estimates $\hat{\theta}_1^{(2)}, \hat{\theta}_2^{(2)}$ and under the first formulation of the problem $\hat{\theta}_3^{(2)}$ which follow distributions

$$\hat{\theta}_1^{(2)} \sim N\left(\theta_1, \frac{4\sigma^2}{\lambda n^{(2)}}\right),$$

$$\hat{\theta}_2^{(2)} \sim N\left(\theta_2, \frac{4\sigma^2}{(1-\lambda)n^{(2)}}\right)$$

and

$$\hat{\theta}_3^{(2)} \sim N\left(\theta_3, \frac{4\sigma^2}{n^{(2)}}\right).$$

These estimates which have the corresponding Z-values

$$Z_1^{(2)} = \frac{\sqrt{\lambda n^{(2)}} \hat{\theta}_1^{(2)}}{2\sigma},$$

$$Z_2^{(2)} = \frac{\sqrt{(1-\lambda)n^{(2)}} \hat{\theta}_2^{(2)}}{2\sigma},$$

and

$$Z_3^{(2)} = \frac{\sqrt{n^{(2)}} \hat{\theta}_3^{(1)}}{2\sigma},$$

and P-values

$$P_1^{(2)} = 1 - \Phi\left(Z_1^{(2)}\right),$$

$$P_2^{(2)} = 1 - \Phi\left(Z_2^{(2)}\right),$$

and

$$P_3^{(2)} = 1 - \Phi\left(Z_3^{(2)}\right).$$

Alternatively the trial may continue only in one sub-population in which case all patients are recruited from this sub-population. Patients are still randomised equally between the new treatment and the control allowing us to estimate the treatment effect. If the trial continues only in sub-population 1 then

$$\hat{\theta}_1^{(2)} \sim N\left(\theta_1, \frac{4\sigma^2}{n^{(2)}}\right),$$

giving the Z-value

$$Z_1^{(2)} = \frac{\sqrt{n^{(2)}} \hat{\theta}_1^{(2)}}{2\sigma},$$

and the P-value

$$P_1^{(2)} = 1 - \Phi\left(Z_1^{(2)}\right).$$

Similarly if the trial continues only in sub-population 2 then

$$\hat{\theta}_2^{(2)} \sim N\left(\theta_2, \frac{4\sigma^2}{n^{(2)}}\right),$$

giving the Z-value

$$Z_2^{(2)} = \frac{\sqrt{n^{(2)}}\hat{\theta}_2^{(2)}}{2\sigma},$$

and the P-value

$$P_2^{(2)} = 1 - \Phi\left(Z_2^{(2)}\right).$$

3.5.2 Analysis

We require strong control of the FWER when conducting our hypothesis tests for Adaptive Enrichment designs. Appropriate testing procedures have been discussed many times for example Bauer and Kohne (1994), Jennison and Turnbull (2007) and Wang et al. (2009), we follow the same methods.

Under the first formulation of the problem we test the null hypotheses $H_{01} : \theta_1 \leq 0$ and $H_{03} : \theta_3 \leq 0$. From the pre-interim recruitment cohort we have P-values $P_1^{(1)}$ and $P_3^{(1)}$ we find a P-value for the intersection hypothesis using Simes method, from Equation 3.4 this is

$$P_{13}^{(1)} = \min(2\min(P_1^{(1)}, P_3^{(1)}), \max(P_1^{(1)}, P_3^{(1)})).$$

From the post-interim cohort the p-values available depend on the interim decision: if we continue to recruit from both sub-populations then we get P-values $P_1^{(2)}$ and $P_3^{(2)}$ and find a P-value for the intersection hypothesis using Simes method as we did for the pre-interim recruitment cohort

$$P_{13}^{(2)} = \min(2\min(P_1^{(2)}, P_3^{(2)}), \max(P_1^{(2)}, P_3^{(2)}));$$

if trial continues only recruiting from the sub-population we only get $P_1^{(2)}$ from the data, we set $P_3^{(2)} = 1$ as this ensures H_{03} cannot be rejected overall and the intersection P-value is given by $P_{13}^{(2)} = P_1^{(2)}$ as we now focus our test on the data we have collected. For the individual P-values we have already defined the corresponding Z-values and we may find the equivalent Z-values for the intersection P-values. We use the weighted inverse normal combination test, Equation 3.5, to find combined Z-values

$$Z_1^{(c)} = w_1 Z_1^{(1)} + w_2 Z_1^{(2)},$$

$$Z_3^{(c)} = w_1 Z_1^{(1)} + w_2 Z_3^{(2)}$$

and

$$Z_{13}^{(c)} = w_1 Z_{13}^{(1)} + w_2 Z_{13}^{(2)},$$

and the corresponding P-values

$$P_1^{(c)} = 1 - \Phi \left(Z_1^{(c)} \right),$$

$$P_3^{(c)} = 1 - \Phi \left(Z_3^{(c)} \right),$$

and

$$P_{13}^{(c)} = 1 - \Phi \left(Z_{13}^{(c)} \right).$$

The P-values $P_1^{(c)}$, $P_3^{(c)}$ and $P_{13}^{(c)}$ are used for the closed testing procedure: rejecting H_{01} globally when $P_1^{(c)} \leq \alpha$ and $P_{13}^{(c)} \leq \alpha$ and rejecting H_{03} globally when $P_3^{(c)} \leq \alpha$ and $P_{13}^{(c)} \leq \alpha$.

We use the same testing methods for the second formulation of the problem. The only difference here is that the trial may be enriched into either population, we take the same approach for the intersection hypothesis whichever sub-population is chosen. If the trial continues only in sub-population 1 we set $P_{12}^{(2)} = P_1^{(2)}$ and if the trial continues only in sub-population 2 we set $P_{12}^{(2)} = P_2^{(2)}$.

3.5.3 Example

In this example we focus on the second formulation of our problem. Using a total sample size $n = 200$ we start by splitting this sample size equally between both sub-populations and both recruitment cohorts. We aim to test the null hypotheses $H_{01} : \theta_1 \leq 0$ and $H_{02} : \theta \leq 0$ and choose $w_1 = w_2 = 1/\sqrt{2}$ for the weighted inverse normal.

Under one particular realisation of this trial we observe: $\bar{x}_1^{(1)} = 15.0$ and $\bar{y}_1^{(1)} = -2.0$ from which we find $\hat{\theta}_1^{(1)} = 17.0$ for the first sub-population; and $\bar{x}_2^{(1)} = 10.7$ and $\bar{y}_2^{(1)} = 2.7$ from which we find $\hat{\theta}_2^{(1)} = 8.0$ for the second sub-population. From this we must decide whether to continue the trial in both populations or in a single population only. To make this decision we will define

a simple decision rule,

$$\text{decision} = \begin{cases} \text{continue in both sub-populations} & \text{if } \hat{\theta}_1^{(1)}, \hat{\theta}_2^{(1)} \geq 7.5 \\ \text{continue only in sub-population 1} & \text{if } \hat{\theta}_1^{(1)} \geq \hat{\theta}_2^{(1)} \text{ and } \hat{\theta}_1^{(1)}, \hat{\theta}_2^{(1)} < 7.5 \\ \text{continue only in sub-population 2} & \text{if } \hat{\theta}_1^{(1)} < \hat{\theta}_2^{(1)} \text{ and } \hat{\theta}_1^{(1)}, \hat{\theta}_2^{(1)} < 7.5. \end{cases} \quad (3.6)$$

Applying this decision rule to our observations we continue the trial in both populations. For the second stage of the trial we observe : $\bar{x}_1^{(2)} = 13.2$ and $\bar{y}_1^{(2)} = 0.7$ from which we find $\hat{\theta}_1^{(2)} = 12.5$ for the first sub-population; and $\bar{x}_2^{(2)} = 9.1$ and $\bar{y}_2^{(2)} = 8.4$ from which we find $\hat{\theta}_2^{(2)} = 0.7$ for the second sub-population.

We now find the combined p-values for the trial to find which hypotheses we may reject globally. From our first stage observations we get p-values $p_1^{(1)} = 0.027$ and $p_2^{(1)} = 0.182$, using Simes method we get the intersection p-value for the first stage $p_{12}^{(1)} = 0.054$. During the second stage, again using Simes method, we get $p_1^{(2)} = 0.078$, $p_2^{(2)} = 0.468$ and $p_{12}^{(2)} = 0.156$. As discussed above we find the combined p-values using the weighted inverse normal giving combined p-values $p_1^{(c)} = 0.009$, $p_2^{(c)} = 0.20$ and $p_{12}^{(c)} = 0.029$. In this particular trial we must accept both null hypotheses, even though we have $p_1^{(c)} < \alpha$ the intersection p-value $p_{12}^{(c)} > \alpha$ and so we are unable to reject the null hypothesis globally.

3.6 An alternative presentation of the summary statistics

Understanding fully how the trial is constructed at a patient level is important when designing an actual trial, however our primary concern will be to optimise the interim decision for the Adaptive Enrichment trial and then compare the overall behaviour of these optimised Adaptive Enrichment trials with the fixed sampling alternatives we defined in Section 3.3. The decision rule we used in Equation 3.6 does not depend on the individual patient results but rather the summary statistics. Similarly when conducting the hypothesis test we only require the summary statistics and knowledge of their distributions. We now take

a simplified approach to these summary statistics rather than considering exact patient numbers for the rest of our work.

We may define the distributions of the summary statistics without directly defining the sample size and variance. The Fisher information about the parameter θ in the case of normally distributed observations is defined to be

$$var(\hat{\theta}) = \frac{1}{\mathcal{I}(\theta)}.$$

For normally distributed data (and approximately for large samples in other cases) the estimate of this parameter is distributed as

$$\hat{\theta} \sim N(\theta, \mathcal{I}(\theta)^{-1}).$$

For a clinical trial with total sample size n , split equally between the two treatments the estimate of the treatment effect we have seen that the estimate of the treatment effect has variance $4\sigma^2/n$. Defining

$$\tilde{\mathcal{I}} = \frac{n}{4\sigma^2}$$

we can describe the distributions of the summary statistics for the trial designs we are interested in terms of $\tilde{\mathcal{I}}^{-1}$ in place of $4\sigma^2/n$ that we used previously. The fixed Enrichment trial recruiting only from sub-population 1 testing $H_{01} : \theta_1 \leq 0$ finds the corresponding summary statistic $\hat{\theta}_1$ which follows the distribution

$$\hat{\theta}_1 \sim N(\theta_1, \tilde{\mathcal{I}}^{-1}).$$

Similarly the fixed Enrichment trial recruiting only from sub-population 2 testing $H_{02} : \theta_2 \leq 0$ finds the corresponding summary statistic $\hat{\theta}_2$ which follows the distribution

$$\hat{\theta}_2 \sim N(\theta_2, \tilde{\mathcal{I}}^{-1}).$$

The fixed sampling testing multiple hypotheses varies depending on which formulation of the problem we are working under. Under both formulations

of the problem we find

$$\hat{\theta}_1 \sim N(\theta_1, (\lambda\tilde{\mathcal{I}})^{-1})$$

and

$$\hat{\theta}_2 \sim N(\theta_2, ((1 - \lambda)\tilde{\mathcal{I}})^{-1}).$$

Under the second formulation of the problem we have null hypotheses $H_{01} : \theta_1 \leq 0$ and $H_{02} : \theta_2 \leq 0$. Under the first formulation we have null hypotheses $H_{01} : \theta_1 \leq 0$ and $H_{03} : \theta_3 \leq 0$ and $\hat{\theta}_3 = \lambda\hat{\theta}_1 + (1 - \lambda)\hat{\theta}_2$ giving the joint distribution,

$$\begin{pmatrix} \hat{\theta}_1 \\ \hat{\theta}_3 \end{pmatrix} \sim N_2 \left(\begin{pmatrix} \theta_1 \\ \theta_3 \end{pmatrix}, \begin{pmatrix} (\lambda\tilde{\mathcal{I}})^{-1} & \tilde{\mathcal{I}}^{-1} \\ \tilde{\mathcal{I}}^{-1} & \tilde{\mathcal{I}}^{-1} \end{pmatrix} \right).$$

Adaptive Enrichment distributions

Finally the distributions for Adaptive Enrichment designs become easier to handle. Keeping λ as the proportion of the population in the first sub-population and defining τ as the proportion of the sample from the pre-interim recruitment cohort we may define the distributions for each stage and population.

Under the second formulation of the problem the pre-interim recruitment cohort has summary statistics $\hat{\theta}_1^{(1)}$ and $\hat{\theta}_2^{(1)}$ which have distributions,

$$\hat{\theta}_1^{(1)} \sim N(\theta_1, (\lambda\tau\tilde{\mathcal{I}})^{-1})$$

and

$$\hat{\theta}_2^{(1)} \sim N(\theta_2, ((1 - \lambda)\tau\tilde{\mathcal{I}})^{-1}).$$

The post interim recruitment cohort may then follow one of three options: If the trial continues in both sub-populations then the summary statistics for the post-interim recruitment cohort $\hat{\theta}_1^{(2)}$ and $\hat{\theta}_2^{(2)}$ are distributed as,

$$\hat{\theta}_1^{(2)} \sim N(\theta_1, (\lambda(1 - \tau)\tilde{\mathcal{I}})^{-1})$$

and

$$\hat{\theta}_2^{(2)} \sim N(\theta_2, ((1 - \lambda)(1 - \tau)\tilde{\mathcal{I}})^{-1});$$

if the trial continues in only a single sub-population then only the summary statistic in the sub-population will be observed and all of the remaining sample will be used in this population, where $i = 1$ or 2 denotes on which population is chosen, we have

$$\hat{\theta}_i^{(2)} \sim N(\theta_i, ((1 - \tau)\tilde{\mathcal{I}})^{-1})$$

Whichever option is chosen we have summary statistics with known distributions and may conduct the analysis of the trial as outlined in Section 3.5.2.

Similarly under the first formulation of the problem we now see our summary statistics from the pre-interim cohort $\hat{\theta}_1^{(1)}$ and $\hat{\theta}_3^{(1)}$ following the multivariate normal distribution,

$$\begin{pmatrix} \hat{\theta}_1^{(1)} \\ \hat{\theta}_3^{(1)} \end{pmatrix} \sim N_2 \left(\begin{pmatrix} \theta_1 \\ \theta_3 \end{pmatrix}, \begin{pmatrix} (\lambda\tau\tilde{\mathcal{I}})^{-1} & (\tau\tilde{\mathcal{I}})^{-1} \\ (\tau\tilde{\mathcal{I}})^{-1} & (\tau\tilde{\mathcal{I}})^{-1} \end{pmatrix} \right).$$

This is equivalent to considering the sub-population of interest and complement separately and then combining these afterwards. This gives

$$\hat{\theta}_1^{(1)} \sim N(\theta_1, (\lambda\tau\tilde{\mathcal{I}})^{-1})$$

in the sub-population of interest and

$$\hat{\theta}_2^{(1)} \sim N(\theta_2, ((1 - \lambda)\tau\tilde{\mathcal{I}})^{-1})$$

in the complement, with $\hat{\theta}_3^{(1)} = \lambda\hat{\theta}_1^{(1)} + (1 - \lambda)\hat{\theta}_2^{(1)}$. The advantage of splitting the distributions in this way is it allows for simulation from 2 independent normals which is simpler and provides more detail if desired at the interim analysis.

In this formulation of the problem the post-interim recruitment cohort only has two options. Continuing in the full population will give summary statistics $\hat{\theta}_1$ and $\hat{\theta}_3$ in the sub and full populations respectively, where these have the multivariate normal distribution given by,

$$\begin{pmatrix} \hat{\theta}_1^{(2)} \\ \hat{\theta}_3^{(2)} \end{pmatrix} \sim N_2 \left(\begin{pmatrix} \theta_1 \\ \theta_3 \end{pmatrix}, \begin{pmatrix} (\lambda(1 - \tau)\tilde{\mathcal{I}})^{-1} & ((1 - \tau)\tilde{\mathcal{I}})^{-1} \\ ((1 - \tau)\tilde{\mathcal{I}})^{-1} & ((1 - \tau)\tilde{\mathcal{I}})^{-1} \end{pmatrix} \right).$$

As with the pre-interim cohort we may equivalently write this as two independent distributions for the sub-population and the complement,

$$\hat{\theta}_1^{(2)} \sim N(\theta_1, (\lambda(1 - \tau)\tilde{\mathcal{I}})^{-1})$$

$$\hat{\theta}_2^{(2)} \sim N(\theta_2, ((1 - \lambda)(1 - \tau)\tilde{\mathcal{I}})^{-1})$$

where $\hat{\theta}_3^{(2)} = \lambda\hat{\theta}_1^{(2)} + (1 - \lambda)\hat{\theta}_2^{(2)}$. Alternatively the trial may continue only in the sub-population of interest in which case the only summary statistic is $\hat{\theta}_1^{(2)}$ which uses the rest of the available information and is distributed as,

$$\hat{\theta}_1^{(2)} \sim N(\theta_1, ((1 - \tau)\tilde{\mathcal{I}})^{-1}).$$

As is the case with formulation 2 whichever option is chosen we have summary statistics with known distributions and may conduct the analysis of the trial as outlined in Section 3.5.2. We shall use these distributions described using $\tilde{\mathcal{I}}$ defined here throughout the examples to follow.

Chapter 4

Evaluating performance and optimisation

4.1 Methods for comparison

When assessing the performance of trials we must understand what good performance of a trial will look like. For Adaptive Enrichment designs we want to compare the overall performance with competing fixed sampling designs, learning where we may see some overall benefit from adaptation particularly when the fixed sampling designs are doing well. The comparisons that may be made will differ depending on the formulation of the problem being considered.

4.1.1 Running example

Throughout this chapter we shall keep returning to the same core example using a fixed set of parameters to define the recruitment of each of the trial designs. Let

$$\tilde{\mathcal{I}} = \left(\frac{\Phi^{-1}(0.9) + \Phi^{-1}(0.975)}{10} \right)^2,$$

for all trials essentially fixing a total sample size, we saw in Section 3.6 how we use $\tilde{\mathcal{I}}$ in place of the sample size and variance. For the designs that make use of sub-populations we take $\lambda = 1/2$, that is the first sub-population accounts for half of the sample. The Adaptive Enrichment designs will also have multiple stages, by taking $\tau = 1/2$ we assume that half of the total sample is used in each

recruitment cohort. With these defined we may use the distributions given in Section 3.6 for each of the possible trial designs.

Each of the trial methods require strong control of the FWER as defined in Equation 3.1. For the designs that test multiple hypotheses we use a closed testing procedure described in Section 3.2.2. For the intersection hypothesis we use Simes method seen in Equation 3.4. To ensure strong control of the FWER in the Adaptive Enrichment trial we follow the analysis method from Section 3.5.2: we apply Simes method within each recruitment cohort finding $P_{13}^{(i)}$ for $i = 1, 2$ under the first formulation of the problem and $P_{12}^{(i)}$ for $i = 1, 2$ under the second formulation of the problem. When a single population is selected for the post interim recruitment cohort the intersection p-value is given by the individual p-value for testing that sub-population. We combine the p-values from each recruitment cohort using the weighted inverse normal combination test, Section 3.4.1 with pre-defined weights $w_1 = \sqrt{\tau}$ and $w_2 = \sqrt{1 - \tau}$. We will use this testing method throughout this chapter under both formulations of the problem setting $\alpha = 0.025$.

4.1.2 Formulation 1

Recall from section 3.1.2 that under the first formulation of the problem we consider a single sub-population of interest within the full population of patients eligible for the trial. This gives two null hypothesis $H_{01} : \theta_1 \leq 0$ in the sub-population of interest and $H_{03} : \theta_3 \leq 0$ in the full population.

When using fixed sampling methods we may conduct the trial in one of three ways: a fixed Enrichment trial where we recruit only from the sub-population of interest and therefore only test $H_{01} : \theta_1 \leq 0$; a fixed sampling trial testing multiple hypotheses where we recruit in both the sub-population of interest and the complement allowing us to test both $H_{01} : \theta_1 \leq 0$ and $H_{03} : \theta_3 \leq 0$; or we could conduct a trial using only the full population and only test $H_{03} : \theta_3 \leq 0$. We will not use this final option in our comparisons as the point of Adaptive Enrichment is to make best use of sub-populations where we place emphasis on the case where $\theta_1 > \theta_2$.

For this formulation we allow an Adaptive Enrichment with a single interim analysis where the options are to continue with both the sub and full populations, or to enrich the trial continuing only in the sub-population of interest. Continuing in both populations will allow testing of both $H_{01} : \theta_1 \leq 0$ and $H_{03} : \theta_3 \leq 0$, while continuing only in the sub-population only allows testing of $H_{01} : \theta_1 \leq 0$.

Example

Using the core example given in Section 4.1.1 we have most of what we require to find basic properties of all three trial designs. The Adaptive Enrichment trial also requires a decision rule. We will come to the optimisation of this rule later in the Chapter, for now we use a simple rule of the form

$$\text{decision} = \begin{cases} \text{continue in both populations} & \text{if } \hat{\theta}_3^{(1)} \geq \psi \\ \text{continue only in the sub-population} & \text{if } \hat{\theta}_3^{(1)} < \psi, \end{cases} \quad (4.1)$$

and set $\psi = 7.5$ for this example. We will return to this decision rule when we begin optimisation of Adaptive Enrichment trials in Section 4.3.4.

We find the operating characteristics of both fixed sampling trials and the Adaptive Enrichment design. There are three outcomes we focus on for each trial: only H_{01} is rejected, denote this event by \mathcal{R}_1 ; both H_{01} and H_{03} are rejected, denote this event by \mathcal{R}_b ; or only H_{03} is rejected, denote this event by \mathcal{R}_3 . The probabilities of these events for each type of trial can be seen in Table 4.1. The probabilities for the fixed sampling designs may be computed directly using the mvtnorm package in R when required, Genz et al. (2008). For the Adaptive Enrichment design we simulated 1,000,000 realisations of the summary statistics of the trials in order to estimate these probabilities with a standard deviation in the worst case of

$$\sqrt{\frac{0.5^2}{10^6}} = 0.0005.$$

In the first section of Table 4.1 we see that the fixed Enrichment design that samples and tests the null hypothesis only in the sub-population performs

Fixed sampling: sub-population only						
θ_1	θ_2	θ_3	$\mathbb{P}(\mathcal{R}_1)$	$\mathbb{P}(\mathcal{R}_3)$	$\mathbb{P}(\mathcal{R}_b)$	$\mathbb{P}(\mathcal{R}_1) + \mathbb{P}(\mathcal{R}_3) + \mathbb{P}(\mathcal{R}_b)$
10	10	10	0.900	-	-	0.900
7.5	7.5	7.5	0.674	-	-	0.674
5	5	5	0.362	-	-	0.362
10	0	5	0.900	-	-	0.900
7.5	0	3.75	0.674	-	-	0.674
5	0	2.5	0.362	-	-	0.362
10	-10	0	0.900	-	-	0.900
Fixed sampling: both populations						
θ_1	θ_2	θ_3	$\mathbb{P}(\mathcal{R}_1)$	$\mathbb{P}(\mathcal{R}_3)$	$\mathbb{P}(\mathcal{R}_b)$	$\mathbb{P}(\mathcal{R}_1) + \mathbb{P}(\mathcal{R}_3) + \mathbb{P}(\mathcal{R}_b)$
10	10	10	0.006	0.238	0.619	0.862
7.5	7.5	7.5	0.014	0.227	0.375	0.617
5	5	5	0.021	0.128	0.164	0.313
10	0	5	0.214	0.016	0.334	0.564
7.5	0	3.75	0.141	0.020	0.185	0.347
5	0	2.5	0.070	0.019	0.084	0.174
10	-10	0	0.495	0.000	0.025	0.520
Adaptive Enrichment						
θ_1	θ_2	θ_3	$\mathbb{P}(\mathcal{R}_1)$	$\mathbb{P}(\mathcal{R}_3)$	$\mathbb{P}(\mathcal{R}_b)$	$\mathbb{P}(\mathcal{R}_1) + \mathbb{P}(\mathcal{R}_3) + \mathbb{P}(\mathcal{R}_b)$
10	10	10	0.164	0.153	0.516	0.833
7.5	7.5	7.5	0.177	0.128	0.275	0.580
5	5	5	0.120	0.065	0.105	0.289
10	0	5	0.501	0.008	0.192	0.701
7.5	0	3.75	0.342	0.010	0.101	0.452
5	0	2.5	0.166	0.009	0.044	0.219
10	-10	0	0.665	0.000	0.012	0.677

Table 4.1: Comparing operating characteristics under the first formulation of the problem

exactly as expected, the probability of rejecting the null hypothesis falls as the true treatment effect falls and is unaffected by the treatment effect in the full population. The second section shows a fixed sampling design where multiple hypotheses are tested. If both treatment effects are strong the overall probability of rejecting a null hypothesis are similar between the two fixed sampling designs, with a good chance of rejecting both null hypotheses or just the null hypothesis in the full population when testing multiple hypotheses. However when the treatment effect is only present in the sub-population the overall probability of rejecting null hypotheses is much lower for the design testing multiple hypotheses when compared to the fixed Enrichment.

Results for the Adaptive Enrichment design, shown in the third section of Table 4.1, fall somewhere between the two fixed sampling methods. When both treatment effects are high the Adaptive Enrichment design allows rejection of both hypotheses with a slightly lower probability in the full population than the fixed design testing multiple hypotheses. When a treatment effect is only present in the sub-population of interest we see that some of the probability of rejecting this hypothesis under the Adaptive Enrichment design lies between the two fixed sampling alternatives.

4.1.3 Formulation 2

The second formulation of our problem also introduced in section 3.1.2 uses two independent sub-populations within the patient population and is not concerned with the treatment effect in the full population. The corresponding null hypotheses are $H_{01} : \theta_1 \leq 0$ in the first sub-population and $H_{02} : \theta_2 \leq 0$ in the second sub-population.

In this formulation of the problem the fixed sampling methods that we make comparisons with are constructed slightly differently. As before we may conduct a fixed Enrichment trial, however this may be in either the first or second sub-population. This means recruiting in a single sub-population and then testing either $H_{01} : \theta_1 \leq 0$ or $H_{02} : \theta_2 \leq 0$ as appropriate. As in the first formulation we may still recruit patients from both sub-populations allowing us to test both $H_{01} : \theta_1 \leq 0$ and $H_{02} : \theta_2 \leq 0$.

As with the first formulation of the problem we allow an Adaptive Enrichment design with a single interim analysis. At this point we may continue the trial in either sub-population 1 or sub-population 2 only allowing us to test $H_{01} : \theta_1 \leq 0$ or $H_{02} : \theta_2 \leq 0$ respectively, or we may continue in both sub-populations testing both $H_{01} : \theta_1 \leq 0$ and $H_{02} : \theta_2 \leq 0$.

Example

Returning to our core example defined in Section 4.1.1 we learn about the competing trial designs. We define a decision rule for the Adaptive Enrichment trial that differs slightly from the first formulation of the problem since the options at the interim analysis have changed. We use a rule of the form

$$\text{decision} = \begin{cases} \text{continue in both sub-populations} & \text{if } \hat{\theta}_1^{(1)} \geq \psi \text{ and } \hat{\theta}_2^{(1)} \geq \psi \\ \text{continue only in sub-population 1} & \text{if } \hat{\theta}_1^{(1)} \geq \hat{\theta}_2^{(1)} \text{ and } \hat{\theta}_2^{(1)} < \psi \\ \text{continue only in sub-population 2} & \text{if } \hat{\theta}_2^{(1)} > \hat{\theta}_1^{(1)} \text{ and } \hat{\theta}_1^{(1)} < \psi, \end{cases} \quad (4.2)$$

where we set $\psi = 7.5$.

The final difference in this formulation is in the possible outcomes of the trial: if only H_{01} is rejected we denote this event by \mathcal{R}_1 , if both H_{01} and H_{02} are rejected, denote this event by \mathcal{R}_b and if only H_{02} is rejected we denote this event by \mathcal{R}_2 . The probabilities of these events for each trial design can be seen in Table 4.2.

Table 4.2 allows us to compare the fixed sampling designs with a simple Adaptive Enrichment design for the second formulation of the problem. We see that when the treatment effect in a sub-population is high the fixed Enrichment trial in this population is performing well for the individual hypothesis. However we see the issue that the fixed Enrichment trial is not robust when the treatment effect is in the opposite sub-population, in this respect both the fixed sample testing multiple hypotheses and the Adaptive Enrichment design offer an improvement. The Adaptive Enrichment design gives a higher power of rejecting at least one null hypotheses than the fixed sample testing multiple hypotheses in all of these examples, this is largely down to the additional probability of rejecting the null hypothesis in the sub-population with the higher treatment effect.

		Fixed sampling: sub-population 1 only			
θ_1	θ_2	$\mathbb{P}(\mathcal{R}_1)$	$\mathbb{P}(\mathcal{R}_2)$	$\mathbb{P}(\mathcal{R}_b)$	$\mathbb{P}(\mathcal{R}_1) + \mathcal{P}(\mathcal{R}_2) + \mathcal{P}(\mathcal{R}_b)$
10	10	0.900	-	-	0.900
5	5	0.362	-	-	0.362
10	5	0.900	-	-	0.900
5	0	0.362	-	-	0.362
5	10	0.362	-	-	0.362
0	5	0.025	-	-	0.025

		Fixed sampling: sub-population 2 only			
θ_1	θ_2	$\mathbb{P}(\mathcal{R}_1)$	$\mathbb{P}(\mathcal{R}_2)$	$\mathbb{P}(\mathcal{R}_b)$	$\mathbb{P}(\mathcal{R}_1) + \mathcal{P}(\mathcal{R}_2) + \mathcal{P}(\mathcal{R}_b)$
10	10	-	0.900	-	0.900
5	5	-	0.362	-	0.362
10	5	-	0.362	-	0.362
5	0	-	0.025	-	0.025
5	10	-	0.900	-	0.900
0	5	-	0.362	-	0.362

		Fixed sampling: both populations			
θ_1	θ_2	$\mathbb{P}(\mathcal{R}_1)$	$\mathbb{P}(\mathcal{R}_2)$	$\mathbb{P}(\mathcal{R}_b)$	$\mathbb{P}(\mathcal{R}_1) + \mathcal{P}(\mathcal{R}_2) + \mathcal{P}(\mathcal{R}_b)$
10	10	0.192	0.192	0.397	0.781
5	5	0.108	0.108	0.043	0.260
10	5	0.412	0.051	0.131	0.594
5	0	0.133	0.010	0.005	0.148
5	10	0.051	0.412	0.131	0.594
0	5	0.010	0.133	0.005	0.148

		Adaptive Enrichment			
θ_1	θ_2	$\mathbb{P}(\mathcal{R}_1)$	$\mathbb{P}(\mathcal{R}_2)$	$\mathbb{P}(\mathcal{R}_b)$	$\mathbb{P}(\mathcal{R}_1) + \mathcal{P}(\mathcal{R}_2) + \mathcal{P}(\mathcal{R}_b)$
10	10	0.284	0.284	0.276	0.844
5	5	0.145	0.145	0.025	0.315
10	5	0.520	0.079	0.083	0.682
5	0	0.179	0.010	0.003	0.192
5	10	0.079	0.520	0.083	0.682
0	5	0.010	0.179	0.003	0.192

Table 4.2: Comparing operating characteristics under the first formulation of the problem

4.2 Constructing a single measure of trial performance

In the examples of section 4.1 we see that each formulation has four mutually exclusive possible outcomes, three of which involve the rejection of at least one null hypotheses. The final columns of Tables 4.1 and 4.2 show the probability of rejecting any combination of null hypotheses, however the overall probability of a positive outcome may not fully capture the relative importance of each possible outcome.

4.2.1 Weighting outcomes

Our first aim is to construct a single measure that accounts for the fact that each hypothesis represents a different population. Let

$$\boldsymbol{\theta} = \begin{cases} (\theta_1, \theta_3) & \text{under the first formulation} \\ (\theta_1, \theta_2) & \text{under the second formulation.} \end{cases}$$

Taking the probability of rejecting any combination of null hypotheses as a starting point, under the first formulation of our problem this is

$$\mathbb{P}_{\boldsymbol{\theta}}(\mathcal{R}_1 \cup \mathcal{R}_3 \cup \mathcal{R}_b) = \mathbb{P}_{\boldsymbol{\theta}}(\mathcal{R}_1) + \mathbb{P}_{\boldsymbol{\theta}}(\mathcal{R}_3) + \mathbb{P}_{\boldsymbol{\theta}}(\mathcal{R}_b).$$

Equivalently under the second formulation we have

$$\mathbb{P}_{\boldsymbol{\theta}}(\mathcal{R}_1 \cup \mathcal{R}_2 \cup \mathcal{R}_b) = \mathbb{P}_{\boldsymbol{\theta}}(\mathcal{R}_1) + \mathbb{P}_{\boldsymbol{\theta}}(\mathcal{R}_2) + \mathbb{P}_{\boldsymbol{\theta}}(\mathcal{R}_b).$$

The issue with this is that equal importance is assigned to each possible outcome of the trial, when in fact the proportion of the population we will claim receive some benefit from the new treatment varies depending on which hypotheses have been rejected.

We define the weighted sum of probabilities of rejecting null hypotheses as

$$\mathcal{W}(\boldsymbol{\theta}) = \omega_1 \mathbb{P}_{\boldsymbol{\theta}}(\mathcal{R}_1) + \omega_3 \mathbb{P}_{\boldsymbol{\theta}}(\mathcal{R}_3) + \omega_b \mathbb{P}_{\boldsymbol{\theta}}(\mathcal{R}_b) \quad (4.3)$$

under the first formulation of the problem and

$$\mathcal{W}(\boldsymbol{\theta}) = \omega_1 \mathbb{P}_{\boldsymbol{\theta}}(\mathcal{R}_1) + \omega_2 \mathbb{P}_{\boldsymbol{\theta}}(\mathcal{R}_2) + \omega_b \mathbb{P}_{\boldsymbol{\theta}}(\mathcal{R}_b). \quad (4.4)$$

under the second formulation of the problem.

Example

Returning to our examples from Section 4.1 we compare the different trial designs using a weighting scheme to see how this may change our interpretation of the results, this will also provide insight into the roles of the choice of weights. We choose weights equal to the proportion of the population that may receive the new treatment based on the rejection of the corresponding null hypothesis, $\omega_1 = \lambda = 1/2$, $\omega_2 = 1 - \lambda = 1/2$, $\omega_3 = 1$ and $\omega_b = 1$.

The top half of Table 4.3 shows comparisons between the competing trial designs for these weights under the first formulation of our problem. We see that the Adaptive Enrichment design is the second best design in most cases, with the fixed sampling design testing multiple hypotheses being the best. However this may reflect that the weighting scheme does not fully reflect which the most positive outcomes are. For example when $\theta_1 = 10$ and $\theta_3 = 5$ we might want to reward the fixed Enrichment design for choosing the sub-population who receive a larger benefit, however our weighting strategy is not capturing this. Since the true treatment effect is not taken into account we only reward the rejection of null hypothesis, of equal importance is whether the new treatment actually provides a benefit to the patients.

The bottom half of Table 4.3 shows similar comparisons for these choices of the weights under the second formulation of our problem, the Adaptive Enrichment trial appears to be performing better than it did under the first formulation. We see that Adaptive Enrichment is always close to the design that maximises $\mathcal{W}(\boldsymbol{\theta})$. Due to the overall reduction in probabilities seen in the fixed design testing multiple hypotheses we do not see that this design is always the best choice under this weighting strategy, however we may still wish to reflect the importance of rejecting different null hypotheses under different combinations of $\boldsymbol{\theta}$.

Formulation 1			$\mathcal{W}(\boldsymbol{\theta})$ under:		
θ_1	θ_2	θ_3	FE ₁	MH	AE
10	10	10	0.450	0.859	0.751
7.5	7.5	7.5	0.377	0.609	0.492
5	5	5	0.181	0.303	0.229
10	0	5	0.450	0.457	0.451
7.5	0	3.75	0.377	0.277	0.281
5	0	2.5	0.181	0.138	0.136
10	-10	0	0.450	0.273	0.273

Formulation 2		$\mathcal{W}(\boldsymbol{\theta})$ under:			
θ_1	θ_2	FE ₁	FE ₂	MH	AE
10	10	0.450	0.450	0.589	0.560
7.5	7.5	0.377	0.377	0.343	0.353
5	5	0.181	0.181	0.151	0.170
10	5	0.450	0.181	0.362	0.383
7.5	2.5	0.377	0.063	0.186	0.221
5	0	0.181	0.013	0.077	0.098

Table 4.3: Comparing trial designs using a weighted sum: FE₁ = Fixed Enrichment in sub-population 1, FE₂ = Fixed Enrichment in sub-population 2, MH = Fixed sampling testing multiple hypotheses, AE = Adaptive Enrichment

4.2.2 Gain functions

The strategy of weighting together the possible outcomes is useful as it gives us a single measure for different combinations of $\boldsymbol{\theta}$ that accounts for the differences in the patient populations that the null hypotheses represent. However this only depends on the rejection of the null hypotheses. We would also like to reflect the true benefit provided to patients; to do this we will construct a utility function to give a higher weight to outcomes where the corresponding treatment effect is also large. An introduction to this that also gives background on Bayesian decision making that we will use later is given by Berger (2013). In particular we focus on defining a gain function as we consider things in terms of the gain associated with particular outcomes of the trial. We could consider this as the gain from the perspective of the company, or the gain for the patients receiving an improved treatment.

We define the gain function in terms of observed outcomes, defining the gain of each possible outcome. If \mathcal{R}_i is true we assign a gain $\gamma_i(\boldsymbol{\theta})$ for each $i = 1, 2, 3, b$ that also reflects the benefit of a true treatment effect of $\boldsymbol{\theta}$. Since the outcomes are mutually exclusive the overall gain of the trial, $G(\boldsymbol{\theta})$ say, will be given by the function $\gamma_i(\boldsymbol{\theta})$ corresponding to the null hypotheses that are rejected. For example under the first formulation of our problem we have

$$G(\boldsymbol{\theta}) = \begin{cases} \gamma_1(\boldsymbol{\theta}) & \text{if } H_{01} \text{ is rejected} \\ \gamma_3(\boldsymbol{\theta}) & \text{if } H_{03} \text{ is rejected} \\ \gamma_b(\boldsymbol{\theta}) & \text{if } H_{01} \text{ and } H_{03} \text{ are rejected.} \end{cases}$$

We assume here that the $\gamma_i(\boldsymbol{\theta})$ depend only on the true treatment effects. In practice the definition might be extended to include things such as the estimates of the treatment effects from the trial.

To capture the gain more clearly we define indicators $\mathbb{I}(\mathcal{R}_i)$ for $i = 1, 2, 3, b$ for the events \mathcal{R}_i as

$$\mathbb{I}(\mathcal{R}_i) = \begin{cases} 1 & \text{if } \mathcal{R}_i \text{ is true} \\ 0 & \text{if } \mathcal{R}_i \text{ is false.} \end{cases}$$

The important point here is that the indicators are about the final outcome of the trial as we do not achieve any gain until the trial is complete. We define κ as the observations from the trial and write the gain function as $G(\boldsymbol{\theta}, \kappa)$ to stress that the outcome of the trial depends on the observed data. We choose κ here since in practice this may represent any information relevant gain of the new treatment, in addition to observation of efficacy (such as we use in our examples) we may also incorporate safety data or even new information about the disease for example if a new treatment becomes available this may change the gain of particular outcomes. We now write $G(\boldsymbol{\theta}, \kappa)$ as the sum of the indicator variables multiplied by the corresponding gain for each possible outcome of the trial. Under the first formulation of our problem the gain is given by

$$G(\boldsymbol{\theta}, \kappa) = \gamma_1(\boldsymbol{\theta})\mathbb{I}(\mathcal{R}_1) + \gamma_3(\boldsymbol{\theta})\mathbb{I}(\mathcal{R}_3) + \gamma_b(\boldsymbol{\theta})\mathbb{I}(\mathcal{R}_b) \quad (4.5)$$

and similarly under the second formulation of our problem we have

$$G(\boldsymbol{\theta}, \kappa) = \gamma_1(\boldsymbol{\theta})\mathbb{I}(\mathcal{R}_1) + \gamma_2(\boldsymbol{\theta})\mathbb{I}(\mathcal{R}_2) + \gamma_b(\boldsymbol{\theta})\mathbb{I}(\mathcal{R}_b). \quad (4.6)$$

Our particular choice for the gain function may not be suitable for all studies, but it need not be rigidly adhered to. The principle of optimising the interim decision of Adaptive Enrichment trials relies only on the definition of a gain function not its particular form, the work that follows demonstrates appropriate methods using one particular choice of the gain function for each formulation of the problem. We feel it should reflect the gain of the trial, whether this is the gain to the patient population or the gain from the company perspective. The choice of gain will influence the optimal design, not just in terms of the interim decision rule but also which type of trial design may be suitable.

Ondra et al. (2016) recently introduced two utility functions appropriate to this setting. One that may capture important outcomes from the perspective of the trial sponsor and another that captures important outcomes from a public health perspective, they then compare fixed sampling trials under these utilities. Our proposed general forms of the gain function focus on the possible trial outcomes and evaluating the gain of each for a particular trial, allowing the construction of appropriate gain functions across a number of trials. In addition to the

comparison of fixed sampling designs we extend the use of the gain function to the optimisation of Adaptive Enrichment trials.

We assess the expected performance of a trial by investigating the expected value of the gain function for choices of $\boldsymbol{\theta}$, $\mathbb{E}_{\boldsymbol{\theta}}(G(\boldsymbol{\theta}, \kappa))$, this is given by

$$\begin{aligned}\mathbb{E}_{\boldsymbol{\theta}}(G(\boldsymbol{\theta}, \kappa)) &= \mathbb{E}_{\boldsymbol{\theta}}(\gamma_1(\boldsymbol{\theta})\mathbb{I}(\mathcal{R}_1) + \gamma_3(\boldsymbol{\theta})\mathbb{I}(\mathcal{R}_3) + \gamma_b(\boldsymbol{\theta})\mathbb{I}(\mathcal{R}_b)) \\ &= \gamma_1(\boldsymbol{\theta})\mathbb{E}_{\boldsymbol{\theta}}(\mathbb{I}(\mathcal{R}_1)) + \gamma_3(\boldsymbol{\theta})\mathbb{E}_{\boldsymbol{\theta}}(\mathbb{I}(\mathcal{R}_3)) + \gamma_b(\boldsymbol{\theta})\mathbb{E}_{\boldsymbol{\theta}}(\mathbb{I}(\mathcal{R}_b)) \\ &= \gamma_1(\boldsymbol{\theta})\mathbb{P}_{\boldsymbol{\theta}}(\mathcal{R}_1) + \gamma_3(\boldsymbol{\theta})\mathbb{P}_{\boldsymbol{\theta}}(\mathcal{R}_3) + \gamma_b(\boldsymbol{\theta})\mathbb{P}_{\boldsymbol{\theta}}(\mathcal{R}_b)\end{aligned}\tag{4.7}$$

under the first formulation, similarly under the second formulation

$$\mathbb{E}_{\boldsymbol{\theta}}(G(\boldsymbol{\theta}, \kappa)) = \gamma_1(\boldsymbol{\theta})\mathbb{P}_{\boldsymbol{\theta}}(\mathcal{R}_1) + \gamma_2(\boldsymbol{\theta})\mathbb{P}_{\boldsymbol{\theta}}(\mathcal{R}_2) + \gamma_b(\boldsymbol{\theta})\mathbb{P}_{\boldsymbol{\theta}}(\mathcal{R}_b).\tag{4.8}$$

This is similar to the weighted sum of probabilities we had before but the weights now depend on the true treatment effects. We use this expected gain in the same way we used the weighted sum to make comparisons of different designs for particular values of $\boldsymbol{\theta}$.

Example

Returning to our core example we can see the difference between the weighted sum and the expected gain. To make this comparison we must first define the gain functions under each formulation of the problem.

Starting with the first formulation of the problem we choose the $\gamma_i(\boldsymbol{\theta})$ to reflect the true treatment effect and the proportion of people treated setting: $\gamma_1(\boldsymbol{\theta}) = \lambda\theta_1$ for the gain of rejecting the null hypothesis only in the sub-population of interest; $\gamma_3(\boldsymbol{\theta}) = \gamma_b(\boldsymbol{\theta}) = \theta_3$ for the gain of rejecting the null hypothesis in the full population. These values for the $\gamma_i(\boldsymbol{\theta})$ give an overall gain function of

$$G(\boldsymbol{\theta}, \kappa) = \lambda\theta_1\mathbb{I}(\mathcal{R}_1) + \theta_3\mathbb{I}(\mathcal{R}_3) + \theta_3\mathbb{I}(\mathcal{R}_b).\tag{4.9}$$

Taking the same approach for the second formulation of the problem we set: $\gamma_1(\boldsymbol{\theta}) = \lambda\theta_1$ for the gain of only rejecting the null hypothesis in the first sub-population; $\gamma_2(\boldsymbol{\theta}) = (1 - \lambda)\theta_2$ for the gain of only rejecting the null hypothesis in

the second sub-population; $\gamma_b(\boldsymbol{\theta}) = \lambda\theta_1 + (1 - \lambda)\theta_2$ for the gain of rejecting the null hypotheses in both populations since this is the average benefit received by patients over both populations. These values for the $\gamma_i(\boldsymbol{\theta})$ give an overall gain function of

$$G(\boldsymbol{\theta}, \kappa) = \lambda\theta_1\mathbb{I}(\mathcal{R}_1) + (1 - \lambda)\theta_2\mathbb{I}(\mathcal{R}_2) + (\lambda\theta_1 + (1 - \lambda)\theta_2)\mathbb{I}(\mathcal{R}_b). \quad (4.10)$$

Table 4.4 compares the weighted sum, Equations 4.3 and 4.4, with weights equal to the proportion of the population represented by each hypotheses with the expected gain given in equations 4.9 and 4.10. Under the first formulation of the problem we see that the gain function has separated the fixed sampling methods showing how they are particularly good when used in the most appropriate scenario but when this is not the case there is potentially a large loss in performance. The fixed design testing multiple hypotheses is now clearly the best fixed design when both the sub-population and the complement receive a benefit from the treatment, and the fixed Enrichment design is better when only the sub-population receives a benefit. The Adaptive Enrichment design performs well when compared to the best fixed sampling design in each situation, so across these few scenarios the adaptive trial appears to be providing the compromise we expected.

Similarly under the second formulation we see an increased separation between the fixed sampling designs. We can see that if both treatment effects are high then we should test both hypotheses, however if the treatment effects are equal but smaller it becomes more difficult to choose between the three fixed sampling designs as they all have similar overall performance. If there is only a treatment effect in one sub-population then the fixed Enrichment design in that sub-population is clearly the best option, however if the opposite sub-population was chosen for the fixed Enrichment trial the expected gain is very low. By comparison the Adaptive Enrichment design regularly performs almost as well as the best fixed design in terms of $\mathbb{E}_{\boldsymbol{\theta}}(G(\boldsymbol{\theta}, \kappa))$ under all of these scenarios, so it appears that the compromise made by the Adaptive Enrichment trial is robust to changes in $\boldsymbol{\theta}$.

			Formulation 1					
			$\mathcal{W}(\boldsymbol{\theta})$ under:			$\mathbb{E}_{\boldsymbol{\theta}}(G(\boldsymbol{\theta}, \kappa))$ under:		
θ_1	θ_2	θ_3	FE ₁	MH	AE	FE ₁	MH	AE
10	10	10	0.450	0.859	0.751	4.50	8.59	7.51
7.5	7.5	7.5	0.377	0.609	0.492	2.83	4.57	3.69
5	5	5	0.181	0.303	0.229	0.91	1.51	1.15
10	0	5	0.450	0.457	0.451	4.50	2.82	3.50
7.5	0	3.75	0.377	0.277	0.281	2.83	1.30	1.70
5	0	2.5	0.181	0.138	0.136	0.91	0.43	0.55
10	-10	0	0.450	0.273	0.273	4.50	2.48	3.33

		Formulation 2							
		$\mathcal{W}(\boldsymbol{\theta})$ under:				$\mathbb{E}_{\boldsymbol{\theta}}(G(\boldsymbol{\theta}, \kappa))$ under:			
θ_1	θ_2	FE ₁	FE ₂	MH	AE	FE ₁	FE ₂	MH	AE
10	10	0.450	0.450	0.589	0.560	4.50	4.50	5.89	5.60
7.5	7.5	0.377	0.377	0.343	0.353	2.83	2.83	2.57	2.65
5	5	0.181	0.181	0.151	0.170	0.91	0.91	0.76	0.85
10	5	0.450	0.181	0.362	0.383	4.50	0.91	2.84	3.42
7.5	2.5	0.377	0.063	0.186	0.221	2.83	0.16	1.15	1.52
5	0	0.181	0.013	0.077	0.098	0.91	0.00	0.33	0.45

Table 4.4: Comparing the weighted sum with the expected gain: FE₁ = Fixed Enrichment in sub-population 1, FE₂ = Fixed Enrichment in sub-population 2, MH = Fixed sampling testing multiple hypotheses, AE = Adaptive Enrichment

Comparing designs across the parameter space

Our parameters are not restricted to the few cases we have so far investigated. We can investigate how the different trial designs behave in terms of the expected gain over the space of $\boldsymbol{\theta}$. This will help us understand where each trial design is performing well and where they are performing poorly, in addition we may directly compare the designs finding which is performing better over the whole parameter space.

Formulation 1

Starting with the first formulation of the problem, we will begin by looking at the fixed sampling methods. Figure 4-1 shows how $\mathbb{E}_{\boldsymbol{\theta}}(G(\boldsymbol{\theta}, \kappa))$ changes with

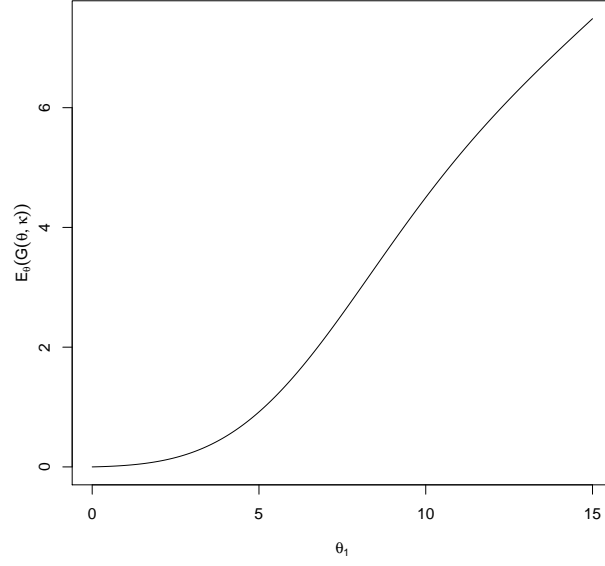


Figure 4-1: $\mathbb{E}_{\boldsymbol{\theta}}(G(\boldsymbol{\theta}, \kappa))$ of fixed Enrichment design

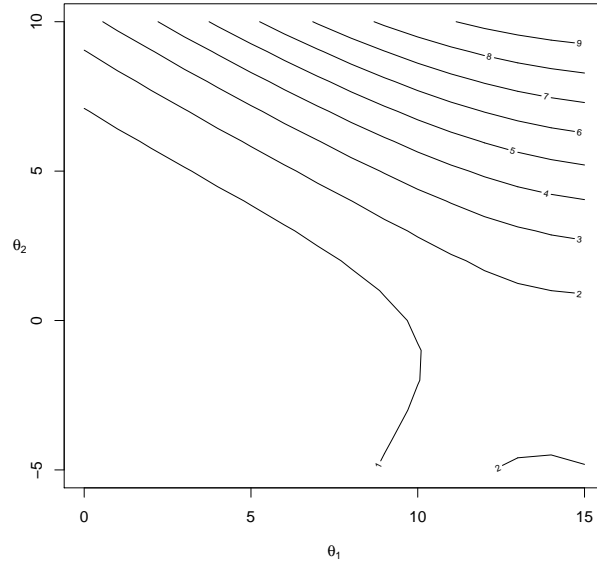


Figure 4-2: $\mathbb{E}_{\boldsymbol{\theta}}(G(\boldsymbol{\theta}, \kappa))$ of the fixed sampling multiple testing design under formulation 1

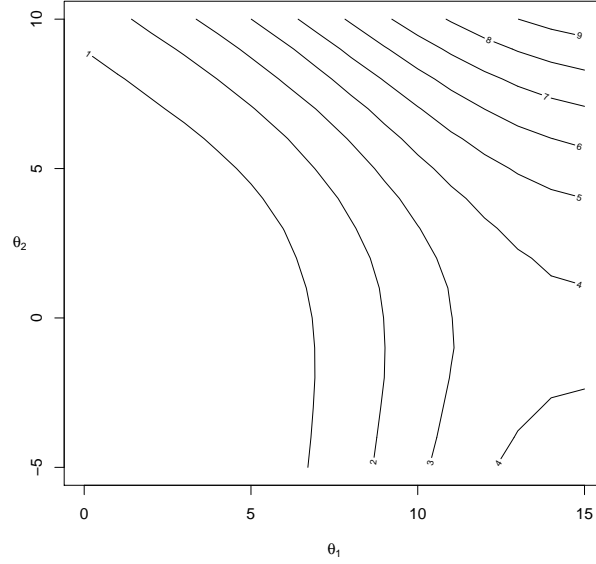


Figure 4-3: $\mathbb{E}_{\theta}(G(\theta, \kappa))$ of the Adaptive Enrichment design under formulation 1

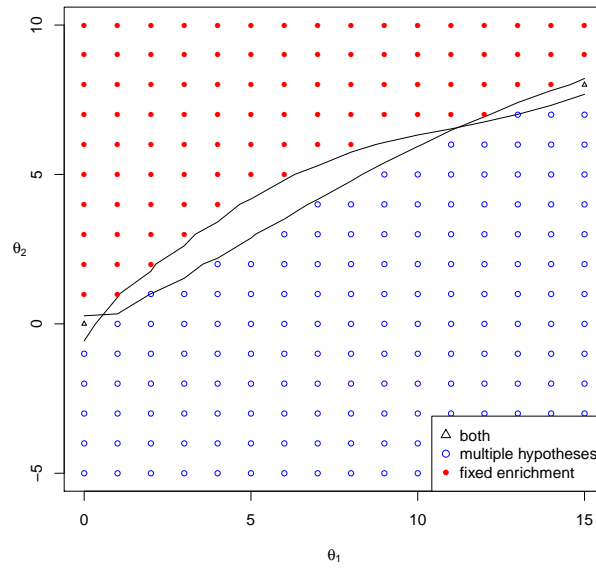


Figure 4-4: Regions where Adaptive Enrichment has higher $\mathbb{E}_{\theta}(G(\theta, \kappa))$ than fixed designs

θ_1 for a fixed Enrichment design in sub-population 1. Unsurprisingly this looks fairly similar to the typical power curve, the plot only shows the varying of θ_1 as the expected gain of the fixed Enrichment design on the treatment effect in the complement of sub-population of interest.

For the fixed sampling design testing multiple hypotheses we must consider how $\mathbb{E}_{\boldsymbol{\theta}}(G(\boldsymbol{\theta}, \kappa))$ varies with both θ_1 and θ_2 . Figure 4-2 shows the contours of $\mathbb{E}_{\boldsymbol{\theta}}(G(\boldsymbol{\theta}, \kappa))$ for θ_1 and θ_2 . Under the fixed Enrichment design these contours would be vertical lines, however when testing multiple hypotheses we see that expected gain changes steeply due to the interaction between θ_1 and θ_2 .

For the Adaptive Enrichment design we again observe how $\mathbb{E}_{\boldsymbol{\theta}}(G(\boldsymbol{\theta}, \kappa))$ varies with both θ_1 and θ_2 . Figure 4-3 shows how the expected gain for the Adaptive Enrichment design varies with the values of θ_1 and θ_2 . The decision rule used when producing this plot is the one given by Equation 4.1 with $\psi = 7.5$. The behaviour of this design is somewhere between the two fixed sampling methods, we still see curvature in the contours due to the interaction between θ_1 and θ_2 however it is not as extreme as the fixed design testing multiple hypotheses and comes closer to vertical when θ_2 is small, as would be expected for the fixed Enrichment design. This again highlights how the Adaptive Enrichment design compromises between the two fixed sampling methods.

We also make a direct comparison of the fixed sampling and Adaptive Enrichment trials, Figure 4-4 shows where in the parameter space the Adaptive Enrichment design performs better than the fixed Enrichment design, the fixed design testing multiple hypotheses or both. We see that when θ_2 is low in relation to θ_1 the Adaptive Enrichment design has a lower expected gain than the fixed Enrichment design but a higher expected gain than the fixed design testing multiple hypotheses. Conversely we see that when θ_2 is high in relation to θ_1 the fixed design testing multiple hypotheses gives a higher expected gain than Adaptive Enrichment and the fixed Enrichment gives a lower expected gain. So as we anticipated in areas where one fixed sampling design is performing poorly the Adaptive Enrichment design takes characteristics similar to the better fixed sampling design and is able to perform better.

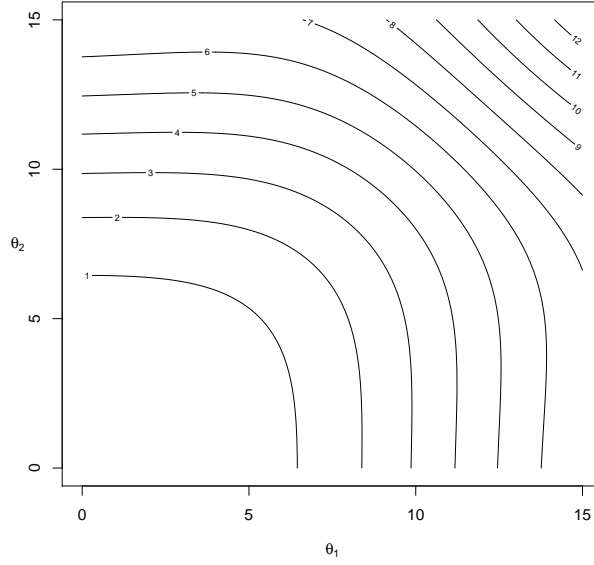


Figure 4-5: $\mathbb{E}_{\theta}(G(\theta, \kappa))$ of the fixed sampling multiple testing design under formulation 2

There is also an area in Figure 4-4 where the Adaptive Enrichment design gives a lower expected gain than both of the fixed sampling alternatives we compare it with. In this region of the plot the fixed sampling designs are providing a similar level of performance to each other, however when the Adaptive Enrichment design chooses to enrich the trial the total sample size in that sub-population is $3/4$ of that given by the fixed Enrichment design. The Adaptive Enrichment design enriching some of the time in this region results in lower probabilities of rejecting null hypotheses and causes the observed reduction in the expected gain, this is a necessary trade off for the ability to make good decisions in more extreme regions of the parameter space.

Formulation 2

Under the second formulation of the problem we have an additional fixed sampling option to compare the Adaptive Enrichment trial to since we may now choose either sub-population for the fixed Enrichment design. Whichever sub-population is chosen for the fixed Enrichment design the trial is recruited only from that sub-population and tests either $H_{01} : \theta_1 \leq 0$ or $H_{02} : \theta_2 \leq 0$.

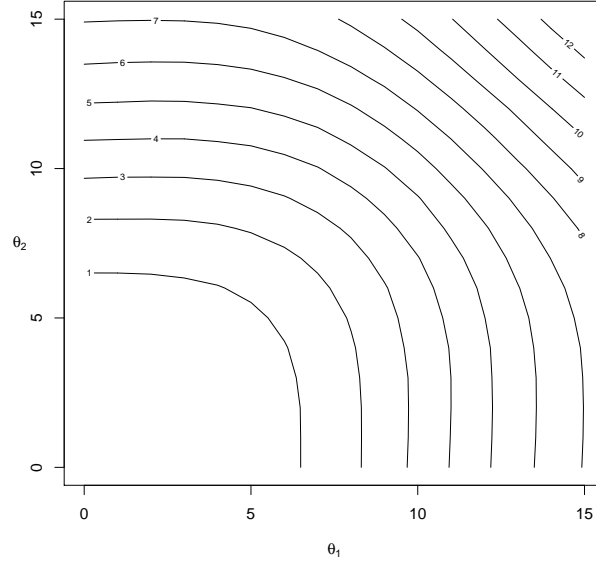


Figure 4-6: $\mathbb{E}_{\boldsymbol{\theta}}(G(\boldsymbol{\theta}, \kappa))$ of the Adaptive Enrichment design under formulation 2

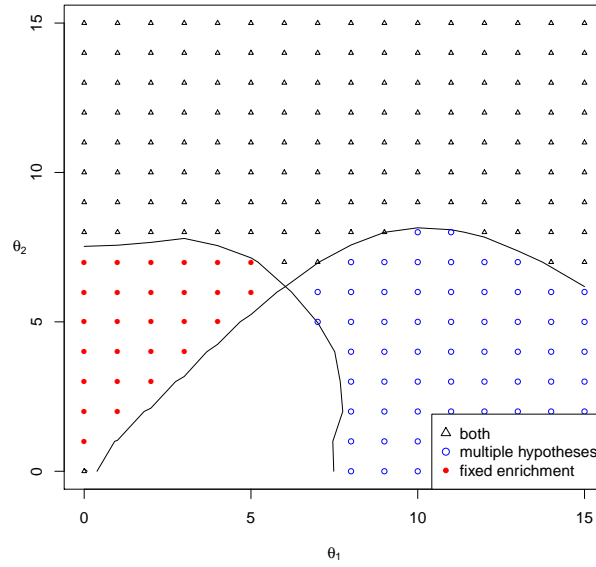


Figure 4-7: Comparing $\mathbb{E}_{\boldsymbol{\theta}}(G(\boldsymbol{\theta}, \kappa))$ of the Adaptive Enrichment design with two of the fixed sampling alternatives under formulation 2

This is equivalent to the fixed Enrichment design we examined under the first formulation, so we may refer back to the results discussed for figure 4-1 since they will be identical regardless of our choice of sub-population.

As before when considering the fixed sampling design testing multiple hypotheses $\mathbb{E}_{\boldsymbol{\theta}}(G(\boldsymbol{\theta}, \kappa))$ will vary with both θ_1 and θ_2 since we aim to test both $H_{01} : \theta_1 \leq 0$ and $H_{02} : \theta_2 \leq 0$. Examining the contours of $\mathbb{E}_{\boldsymbol{\theta}}(G(\boldsymbol{\theta}, \kappa))$ in figure 4-5 we see that the expected gain increases with both θ_1 and θ_2 . The expected gain is now symmetric about the line $\theta_1 = \theta_2$, due to the symmetry of the two sub-populations in this formulation of the problem.

For the Adaptive Enrichment trial we use the decision rule given by equation 4.2 with $\psi = 7.5$, Figure 4-6 shows the contours of $\mathbb{E}_{\boldsymbol{\theta}}(G(\boldsymbol{\theta}, \kappa))$. Under this formulation of the problem the contour given by the Adaptive Enrichment design and the fixed sampling design testing multiple hypotheses take a similar shape, this is because under this formulation of the problem we have $\mathbb{P}_{\boldsymbol{\theta}}(\mathcal{R}_1)$ and $\mathbb{P}_{\boldsymbol{\theta}}(\mathcal{R}_2)$ driving a large part of the expected gain and while the Adaptive Enrichment should improve this as we move further away from $\theta_1 = \theta_2$ this increases slowly and so does not dramatically change the shape of the expected gain.

To fully understand how Adaptive Enrichment is performing across different combinations of θ_1 and θ_2 we show in Figure 4-7 where the Adaptive Enrichment design gives a higher value of $\mathbb{E}_{\boldsymbol{\theta}}(G(\boldsymbol{\theta}, \kappa))$ than the fixed Enrichment trial in sub-population 1, the fixed sampling trial testing multiple hypotheses or both. In the top part of the plot where θ_2 is high we see that the Adaptive Enrichment design is better than both of the fixed sampling designs. When θ_2 is low we see that: Adaptive Enrichment provides an improvement over fixed Enrichment approximately when $\theta_1 > \theta_2$; Adaptive Enrichment provides an improvement over the fixed sampling design testing multiple hypotheses when θ_1 is high. Like the first formulation of the problem there is a region where the Adaptive Enrichment design is worse than both of these fixed sampling options, again this is because when the Adaptive Enrichment design is enriched at the interim analysis it only recruits 3/4 of the sample size that is used in the fixed Enrichment design.

Noting that the fixed Enrichment design in sub-population 2 is symmetric to sub-population 1 (again about the line $\theta_1 = \theta_2$) in this example we see that the region where the Adaptive Enrichment design is worse than all three fixed sampling designs becomes very small. Furthermore we see that there are large regions of the parameter space where the Adaptive Enrichment design is better than two of the three fixed sampling alternatives and when both θ_1 and θ_2 are high it is better than all three even in this unoptimised form.

4.3 Optimisation of adaptive enrichment trials

4.3.1 Oracle decision rule

Knowing that the expected gain is the overall measure of trial performance we can think more carefully about the interim decision to be made in the Adaptive Enrichment trial. Taking $\boldsymbol{\theta}$ as known we can make our first version of an optimal decision. Let κ_1 be the data from the pre-interim recruitment cohort available at the interim analysis and consider the expected gain given κ_1 , $\mathbb{E}_{\boldsymbol{\theta}, \kappa_1}(G(\boldsymbol{\theta}, \kappa))$. Under the first formulation of the problem from Equation 4.8 we get

$$\mathbb{E}_{\boldsymbol{\theta}, \kappa_1}(G(\boldsymbol{\theta}, \kappa)) = \gamma_1(\boldsymbol{\theta})\mathbb{P}_{\boldsymbol{\theta}, \kappa_1}(\mathcal{R}_1) + \gamma_3(\boldsymbol{\theta})\mathbb{P}_{\boldsymbol{\theta}, \kappa_1}(\mathcal{R}_3) + \gamma_b(\boldsymbol{\theta})\mathbb{P}_{\boldsymbol{\theta}, \kappa_1}(\mathcal{R}_b) \quad (4.11)$$

where the probabilities of rejecting null hypotheses now depend on the data that have been observed at the interim analysis. Under the second formulation of the problem from Equation 4.8 we get

$$\mathbb{E}_{\boldsymbol{\theta}, \kappa_1}(G(\boldsymbol{\theta}, \kappa)) = \gamma_1(\boldsymbol{\theta})\mathbb{P}_{\boldsymbol{\theta}, \kappa_1}(\mathcal{R}_1) + \gamma_2(\boldsymbol{\theta})\mathbb{P}_{\boldsymbol{\theta}, \kappa_1}(\mathcal{R}_2) + \gamma_b(\boldsymbol{\theta})\mathbb{P}_{\boldsymbol{\theta}, \kappa_1}(\mathcal{R}_b). \quad (4.12)$$

To evaluate the expected gain at the interim analysis we must find $\mathbb{P}_{\boldsymbol{\theta}, \kappa_1}(\mathcal{R}_i)$ for $i = 1, 2, 3, b$, given the observations, κ_1 , from the pre-interim recruitment cohort. Given the summary statistics from the pre-interim recruitment cohort we can find these probabilities under every possible decision. Thus we may evaluate the expected gain for the rest of the trial for each possible decision at the interim analysis, we choose the decision that maximises the expected gain for the remainder of the trial, this is clearly the optimal choice.

In general we will not know the true value of $\boldsymbol{\theta}$. In this Section we shall suppose that we do know $\boldsymbol{\theta}$ and find the corresponding optimal decision rule, we refer to this as the Oracle decision rule. This procedure provides an upper bound for the performance of procedures that have to rely on an estimate, or a posterior distribution, of $\boldsymbol{\theta}$ at the interim analysis. The Oracle decision rule allows us to clearly see whether Adaptive Enrichment design will be a useful tool for conducting trials in any particular scenario, as we know it is the best possible version of Adaptive Enrichment.

Computation under formulation 1

In order to evaluate $\mathbb{E}_{\boldsymbol{\theta}, \kappa_1}(G(\boldsymbol{\theta}, \kappa))$ we must find $\mathbb{P}_{\boldsymbol{\theta}, \kappa_1}(\mathcal{R}_1)$, $\mathbb{P}_{\boldsymbol{\theta}, \kappa_1}(\mathcal{R}_3)$ and $\mathbb{P}_{\boldsymbol{\theta}, \kappa_1}(\mathcal{R}_b)$. The rejection regions for the null hypotheses based on the summary statistics for the post-interim recruitment cohort are known given κ_1 . Following the testing structure defined in section 4.1.1 we know that applying the closed testing principle,

$$\begin{aligned}\mathbb{P}_{\boldsymbol{\theta}, \kappa_1}(\mathcal{R}_1) &= \mathbb{P}_{\boldsymbol{\theta}, \kappa_1}(Z_1^{(c)} > \Phi^{-1}(1 - \alpha) \cap Z_3^{(c)} \leq \Phi^{-1}(1 - \alpha) \cap Z_{13}^{(c)} > \Phi^{-1}(1 - \alpha)) \\ \mathbb{P}_{\boldsymbol{\theta}, \kappa_1}(\mathcal{R}_3) &= \mathbb{P}_{\boldsymbol{\theta}, \kappa_1}(Z_1^{(c)} \leq \Phi^{-1}(1 - \alpha) \cap Z_3^{(c)} > \Phi^{-1}(1 - \alpha) \cap Z_{13}^{(c)} > \Phi^{-1}(1 - \alpha)) \\ \mathbb{P}_{\boldsymbol{\theta}, \kappa_1}(\mathcal{R}_b) &= \mathbb{P}_{\boldsymbol{\theta}, \kappa_1}(Z_1^{(c)} > \Phi^{-1}(1 - \alpha) \cap Z_3^{(c)} > \Phi^{-1}(1 - \alpha) \cap Z_{13}^{(c)} > \Phi^{-1}(1 - \alpha)).\end{aligned}\tag{4.13}$$

Given the pre-interim recruitment cohort data we know part of $Z_1^{(c)}$, $Z_3^{(c)}$ and $Z_{13}^{(c)}$ when applying the weighted inverse normal combination test, so we may re-write these critical values in terms of the post-interim Z-values $Z_i^{(2)}$ for example

$$\begin{aligned}Z_1^{(c)} &> \Phi^{-1}(1 - \alpha) \\ w_1 Z_1^{(1)} + w_2 Z_1^{(2)} &> \Phi^{-1}(1 - \alpha) \\ Z_1^{(2)} &> \frac{\Phi^{-1}(1 - \alpha) - w_1 Z_1^{(1)}}{w_2}.\end{aligned}$$

Setting the critical values for each $Z_i^{(2)}$ gives critical values $K_{i, \alpha}$, where

$$K_{i, \alpha} = \frac{\Phi^{-1}(1 - \alpha) - w_1 Z_1^{(1)}}{w_2}.$$

So given the pre-interim observations κ_1 if the trial continues in both sub-populations equation 4.13 becomes

$$\begin{aligned}\mathbb{P}_{\boldsymbol{\theta}, \kappa_1}(\mathcal{R}_1) &= \mathbb{P}_{\boldsymbol{\theta}, \kappa_1}(Z_1^{(2)} > K_{1,\alpha} \cap Z_3^{(2)} \leq K_{3,\alpha} \cap Z_{13}^{(2)} > K_{13,\alpha}) \\ \mathbb{P}_{\boldsymbol{\theta}, \kappa_1}(\mathcal{R}_3) &= \mathbb{P}_{\boldsymbol{\theta}, \kappa_1}(Z_1^{(2)} \leq K_{1,\alpha} \cap Z_3^{(2)} > K_{3,\alpha} \cap Z_{13}^{(2)} > K_{13,\alpha}) \\ \mathbb{P}_{\boldsymbol{\theta}, \kappa_1}(\mathcal{R}_b) &= \mathbb{P}_{\boldsymbol{\theta}, \kappa_1}(Z_1^{(2)} > K_{1,\alpha} \cap Z_3^{(2)} > K_{3,\alpha} \cap Z_{13}^{(2)} > K_{13,\alpha}).\end{aligned}\tag{4.14}$$

These probabilities can be made even more precise with further consideration of the intersection Z-value, since we use Simes rule the following are equivalent

$$\begin{aligned}Z_{13}^{(2)} &> K_{13,\alpha} \\ (Z_1^{(2)} > K_{13,\alpha/2} \cup Z_3^{(2)} > K_{13,\alpha/2}) &\cup (Z_1^{(2)} > K_{13,\alpha} \cap Z_3^{(2)} > K_{13,\alpha})\end{aligned}$$

which we use to re-write equation 4.14 only using $Z_1^{(2)}$ and $Z_3^{(2)}$

$$\begin{aligned}\mathbb{P}_{\boldsymbol{\theta}, \kappa_1}(\mathcal{R}_1) &= \mathbb{P}_{\boldsymbol{\theta}, \kappa_1}(Z_1^{(2)} > \max(K_{1,\alpha}, K_{13,\alpha/2}) \cap Z_3^{(2)} \leq \min(K_{3,\alpha}, K_{13,\alpha})) \\ &+ \mathbb{P}_{\boldsymbol{\theta}, \kappa_1}(Z_1^{(2)} > K_{1,\alpha} \cap K_{13,\alpha/2} < Z_3^{(2)} \leq K_{3,\alpha}) \\ &+ \mathbb{P}_{\boldsymbol{\theta}, \kappa_1}(Z_1^{(2)} > \max(K_{1,\alpha}, K_{13,\alpha}) \cap K_{13,\alpha} < Z_3^{(2)} \leq K_{3,\alpha})\end{aligned}\tag{4.15}$$

$$\begin{aligned}\mathbb{P}_{\boldsymbol{\theta}, \kappa_1}(\mathcal{R}_3) &= \mathbb{P}_{\boldsymbol{\theta}, \kappa_1}(Z_1^{(2)} \leq \min(K_{1,\alpha}, K_{13,\alpha}) \cap Z_3^{(2)} > \max(K_{3,\alpha}, K_{13,\alpha/2})) \\ &+ \mathbb{P}_{\boldsymbol{\theta}, \kappa_1}(K_{13,\alpha/2} < Z_1^{(2)} \leq K_{1,\alpha} \cap Z_3^{(2)} > K_{3,\alpha}) \\ &+ \mathbb{P}_{\boldsymbol{\theta}, \kappa_1}(K_{13,\alpha} < Z_1^{(2)} \leq K_{1,\alpha} \cap K_{13,\alpha} < Z_3^{(2)} > \max(K_{3,\alpha}, K_{13,\alpha}))\end{aligned}\tag{4.16}$$

$$\begin{aligned}\mathbb{P}_{\boldsymbol{\theta}, \kappa_1}(\mathcal{R}_b) &= \mathbb{P}_{\boldsymbol{\theta}, \kappa_1}(Z_1^{(2)} > \max(K_{1,\alpha}, K_{13,\alpha/2}) \cap K_{3,\alpha} < Z_3^{(2)} \leq K_{13,\alpha/2}) \\ &+ \mathbb{P}_{\boldsymbol{\theta}, \kappa_1}(K_{1,\alpha} < Z_1^{(2)} \leq K_{13,\alpha/2} \cap Z_3^{(2)} > \max(K_{3,\alpha}, K_{13,\alpha/2})) \\ &+ \mathbb{P}_{\boldsymbol{\theta}, \kappa_1}(Z_1^{(2)} > \max(K_{1,\alpha}, K_{13,\alpha}) \cap Z_3^{(2)} > \max(K_{3,\alpha}, K_{13,\alpha})).\end{aligned}\tag{4.17}$$

We know that if we continue in both sub-populations $Z_1^{(2)} = \lambda(1 - \tau)\tilde{\mathcal{I}}\hat{\theta}_1^{(2)}$ and $Z_3^{(2)} = (1 - \tau)\tilde{\mathcal{I}}\hat{\theta}_3^{(2)}$ and from Section 3.6 we know that $\hat{\theta}_1^{(2)}$ and $\hat{\theta}_3^{(2)}$ have the joint distribution

$$\begin{pmatrix} \hat{\theta}_1^{(2)} \\ \hat{\theta}_3^{(2)} \end{pmatrix} \sim N_2 \left(\begin{pmatrix} \theta_1 \\ \theta_3 \end{pmatrix}, \begin{pmatrix} (\lambda(1 - \tau)\tilde{\mathcal{I}})^{-1} & ((1 - \tau)\tilde{\mathcal{I}})^{-1} \\ ((1 - \tau)\tilde{\mathcal{I}})^{-1} & ((1 - \tau)\tilde{\mathcal{I}})^{-1} \end{pmatrix} \right).$$

and so these probabilities may be found from this multivariate normal distribution.

If we continue only in the sub-population of interest things are simpler as we only need $\mathbb{P}_{\boldsymbol{\theta}, \kappa_1}(\mathcal{R}_1)$ which following the above method may be written as

$$\mathbb{P}_{\boldsymbol{\theta}, \kappa_1}(\mathcal{R}_1) = \mathbb{P}_{\boldsymbol{\theta}, \kappa_1}(Z_1^{(2)} > \max(K_{1,\alpha}, K_{13,\alpha/2})).$$

Where $Z_1^{(2)} = (1 - \tau)\tilde{\mathcal{I}}\hat{\theta}_1^{(2)}$ and from Section 3.6 we know that $\hat{\theta}_1^{(2)}$ follows the distribution

$$\hat{\theta}_1^{(2)} \sim N(\theta_1, ((1 - \tau)\tilde{\mathcal{I}})^{-1})$$

so we may find the probability $\mathbb{P}_{\boldsymbol{\theta}, \kappa_1}(\mathcal{R}_1)$ directly.

Computation under formulation 2

We will not repeat the full derivation under the second formulation as it is mostly the same, the only difference is that at the stage of Equation 4.15, 4.16 and 4.17 we may further split these probabilities using our assumption of independence between the two sub-populations. Thus if we continue in both sub-populations we have

$$\begin{aligned} \mathbb{P}_{\boldsymbol{\theta}, \kappa_1}(\mathcal{R}_1) = & \mathbb{P}_{\boldsymbol{\theta}, \kappa_1}(Z_1^{(2)} > \max(K_{1,\alpha}, K_{12,\alpha/2}))\mathbb{P}_{\boldsymbol{\theta}, \kappa_1}(Z_2^{(2)} \leq \min(K_{2,\alpha}, K_{12,\alpha})) \\ & + \mathbb{P}_{\boldsymbol{\theta}, \kappa_1}(Z_1^{(2)} > K_{1,\alpha})\mathbb{P}_{\boldsymbol{\theta}, \kappa_1}(K_{12,\alpha/2} < Z_2^{(2)} \leq K_{2,\alpha}) \\ & + \mathbb{P}_{\boldsymbol{\theta}, \kappa_1}(Z_1^{(2)} > \max(K_{1,\alpha}, K_{12,\alpha}))\mathbb{P}_{\boldsymbol{\theta}, \kappa_1}(K_{12,\alpha} < Z_2^{(2)} \leq K_{2,\alpha}) \end{aligned} \quad (4.18)$$

$$\begin{aligned} \mathbb{P}_{\boldsymbol{\theta}, \kappa_1}(\mathcal{R}_2) = & \mathbb{P}_{\boldsymbol{\theta}, \kappa_1}(Z_1^{(2)} \leq \min(K_{1,\alpha}, K_{12,\alpha}))\mathbb{P}_{\boldsymbol{\theta}, \kappa_1}(Z_2^{(2)} > \max(K_{2,\alpha}, K_{12,\alpha/2})) \\ & + \mathbb{P}_{\boldsymbol{\theta}, \kappa_1}(K_{12,\alpha/2} < Z_1^{(2)} \leq K_{1,\alpha})\mathbb{P}_{\boldsymbol{\theta}, \kappa_1}(Z_2^{(2)} > K_{2,\alpha}) \\ & + \mathbb{P}_{\boldsymbol{\theta}, \kappa_1}(K_{12,\alpha} < Z_1^{(2)} \leq K_{1,\alpha})\mathbb{P}_{\boldsymbol{\theta}, \kappa_1}(K_{12,\alpha} < Z_2^{(2)} > \max(K_{2,\alpha}, K_{12,\alpha})) \end{aligned} \quad (4.19)$$

$$\begin{aligned} \mathbb{P}_{\boldsymbol{\theta}, \kappa_1}(\mathcal{R}_b) = & \mathbb{P}_{\boldsymbol{\theta}, \kappa_1}(Z_1^{(2)} > \max(K_{1,\alpha}, K_{13,\alpha/2}))\mathbb{P}_{\boldsymbol{\theta}, \kappa_1}(K_{2,\alpha} < Z_2^{(2)} \leq K_{12,\alpha/2}) \\ & + \mathbb{P}_{\boldsymbol{\theta}, \kappa_1}(K_{1,\alpha} < Z_1^{(2)} \leq K_{12,\alpha/2})\mathbb{P}_{\boldsymbol{\theta}, \kappa_1}(Z_2^{(2)} > \max(K_{2,\alpha}, K_{12,\alpha/2})) \quad (4.20) \\ & + \mathbb{P}_{\boldsymbol{\theta}, \kappa_1}(Z_1^{(2)} > \max(K_{1,\alpha}, K_{12,\alpha}))\mathbb{P}_{\boldsymbol{\theta}, \kappa_1}(Z_2^{(2)} > \max(K_{2,\alpha}, K_{12,\alpha})). \end{aligned}$$

If the trial is enriched into either sub-population we have

$$\mathbb{P}_{\boldsymbol{\theta}, \kappa_1}(\mathcal{R}_1) = \mathbb{P}_{\boldsymbol{\theta}, \kappa_1}(Z_1^{(2)} > \max(K_{1,\alpha}, K_{13,\alpha/2})) \quad (4.21)$$

in the first sub-population or

$$\mathbb{P}_{\boldsymbol{\theta}, \kappa_1}(\mathcal{R}_2) = \mathbb{P}_{\boldsymbol{\theta}, \kappa_1}(Z_2^{(2)} > \max(K_{2,\alpha}, K_{12,\alpha/2})) \quad (4.22)$$

in the second sub-population. Again these probabilities may be re-written in terms of $\hat{\theta}_1^{(2)}$ and $\hat{\theta}_2^{(2)}$ which have known distributions from Section 3.6. If the trial continues in both sub-populations we have $Z_1^{(2)} = \lambda(1 - \tau)\tilde{\mathcal{I}}\hat{\theta}_1^{(2)}$ and $Z_2^{(2)} = (1 - \lambda)(1 - \tau)\tilde{\mathcal{I}}\hat{\theta}_2^{(2)}$ where the distributions of $\hat{\theta}_1^{(2)}$ and $\hat{\theta}_2^{(2)}$ are given by

$$\hat{\theta}_1^{(2)} \sim N(\theta_1, (\lambda(1 - \tau)\tilde{\mathcal{I}})^{-1})$$

$$\hat{\theta}_2^{(2)} \sim N(\theta_2, ((1 - \lambda)(1 - \tau)\tilde{\mathcal{I}})^{-1}).$$

If the trial enriches, continuing in only sub-population i ($i = 1, 2$) then $Z_i^{(2)} = (1 - \tau)\tilde{\mathcal{I}}\hat{\theta}_i^{(2)}$, where the distribution for $\hat{\theta}_i^{(2)}$ is given by

$$\hat{\theta}_i^{(2)} \sim N(\theta_i, ((1 - \tau)\tilde{\mathcal{I}})^{-1}).$$

4.3.2 Plotting optimal decision boundaries for Adaptive Enrichment trials

The optimal decision boundaries computed using the conditional expected gain for the interim decision of the Adaptive Enrichment designs do not have a simple form, in order to plot and use them we construct a look up table using a simple sub-dividing algorithm created for this problem. We start by computing the optimal decision for the four corners of the overall grid we wish to compute and iterate the following to find as fine a grid as is required:

- add midpoints between all current grid points
- for each midpoint check the surrounding grid values, if they match assign this value to the new point
- for the midpoints with no clear value compute the optimal decision

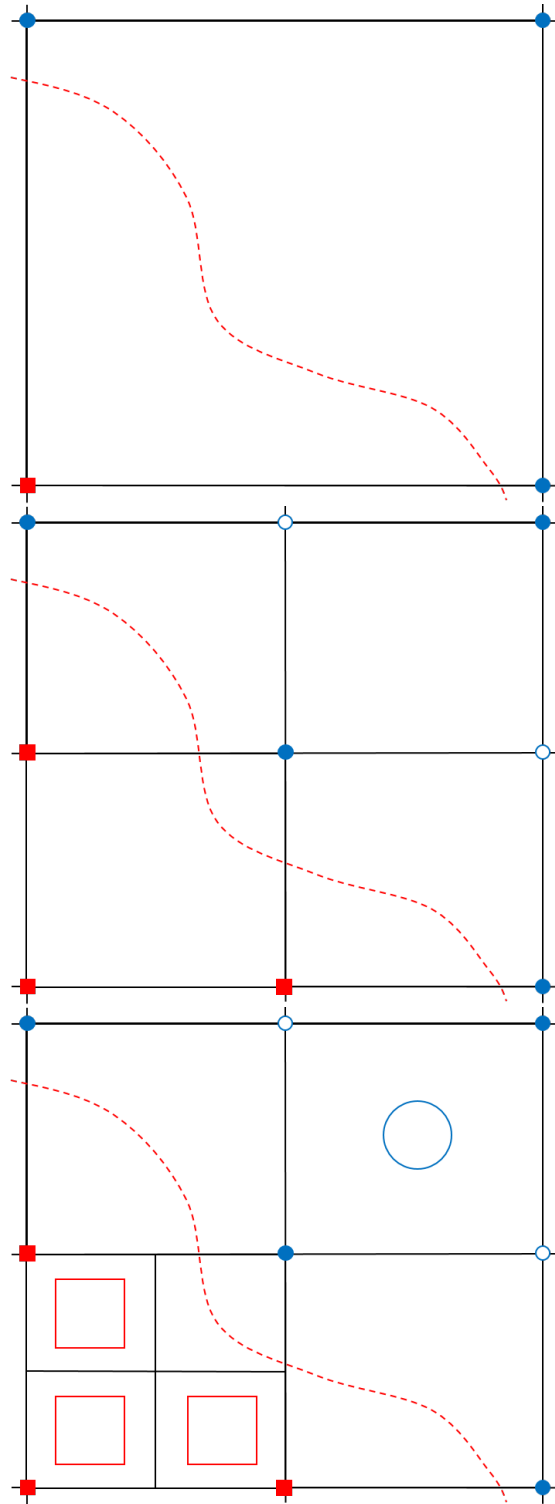


Figure 4-8: Creating a table to apply the optimal decision rule

- merge all points into a new grid.

We can easily illustrate this method in Figure 4-8 our aim is to find the decision boundary given by the dashed line. In the first part of Figure 4-8 we find the optimal decision in each corner of our grid. Adding the midpoints in the second part of Figure 4-8 we compute the optimal decision at the solid symbols and assign the decision at the two white circles. In the third part of Figure 4-8 we see how as we iterate this process we will not need to compute the optimal decision at all of the grid points, in the areas indicated by the white symbols we allocate the existing decision as we continue to find finer grids.

Using this method allows us to find a table giving the optimal decision rules while only computing the optimal decision for a fraction of the grid points. When plotting optimal decision boundaries we create the next level of the grid and plot the points where we would have to compute the optimal decision. To apply an approximate version of the optimal rule we use the decision of the nearest grid point.

Example

Returning to our ongoing example we can now determine the Oracle decision rule for any particular choice of θ . Choosing $\theta_1 = 10$ and $\theta_2 = 5$ Figure 4-9 shows the oracle decision under the first formulation of the problem for a small selection of first stage treatment effect estimates $\hat{\theta}_1^{(1)}$ and $\hat{\theta}_3^{(1)}$. We immediately see how this differs from the simple decision rule we have been using which would be a simple horizontal line through $\hat{\theta}_3^{(1)} = 7.5$.

Similarly under the second formulation of the problem we may find the Oracle decision rule for any particular choice of θ_1 and θ_2 . Figure 4-10 shows the Oracle decision rule under the second formulation of the problem where $\theta_1 = \theta_2 = 10$. As with the first formulation we see that this rule greatly differs from the simple rule we had been using. When the combination of $\hat{\theta}_1^{(1)}$ and $\hat{\theta}_2^{(1)}$ is sufficiently large we continue the trial in both sub-populations, otherwise we continue only in one. There is a blank area where the expected gain is the same for enriching the trial into either sub-population allowing us to choose either, this is because

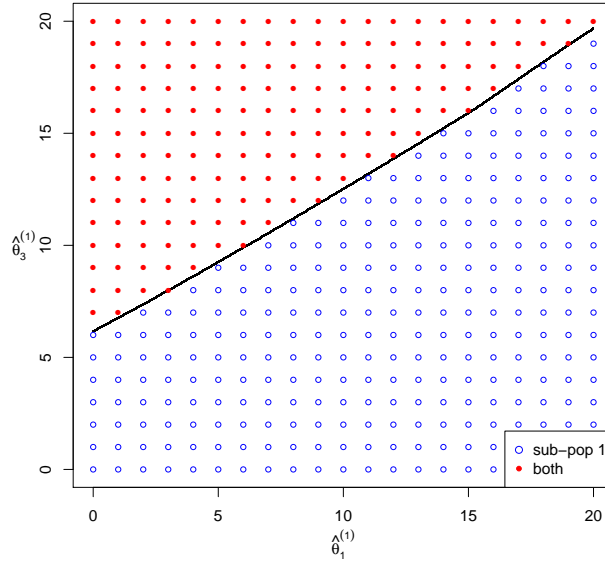


Figure 4-9: Oracle decision rule for $\theta_1 = 10$ and $\theta_3 = 5$ under formulation 1

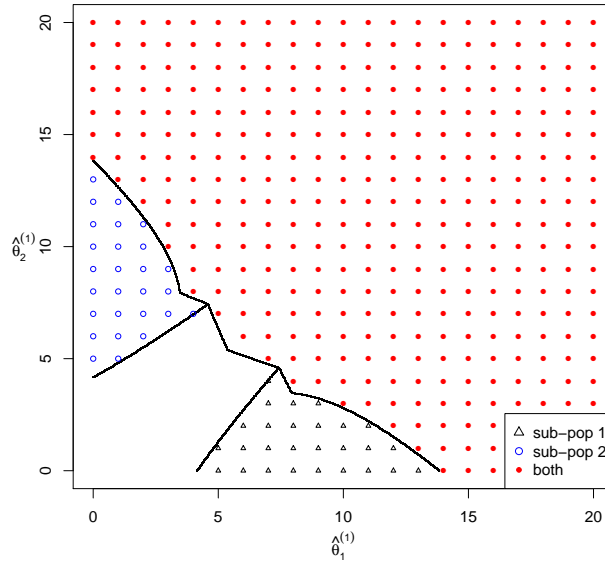


Figure 4-10: Oracle decision rule for $\theta_1 = 10$ and $\theta_3 = 10$ under formulation 2

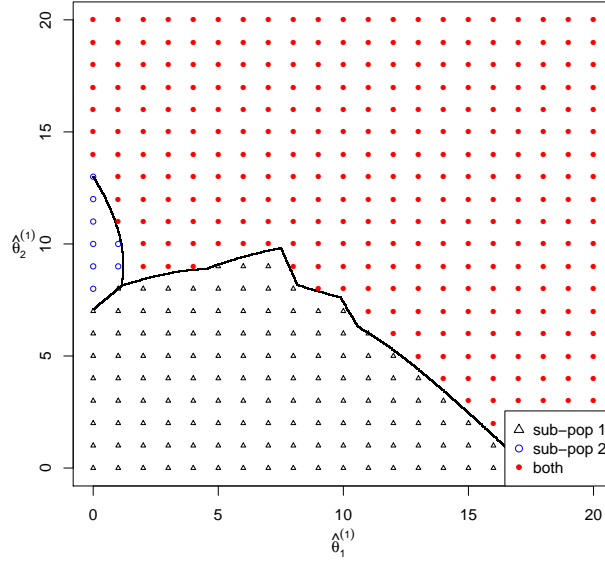


Figure 4-11: Oracle decision rule for $\theta_1 = 10$ and $\theta_3 = 8$ under formulation 2

the first stage p-values are giving $P_{12}^{(1)} = \max(P_1^{(1)}, P_2^{(1)})$ so to reject either null hypothesis globally in stage 2 the critical region is defined by $P_{12}^{(1)}$ and since the treatment effects are the same the probability of rejecting each null hypothesis is equal when enriching into either sub-population. The discontinuities in the boundaries between continuing in both sub-populations and continuing in a single sub-population are a consequence of using Simes method for the intersection p-value within each stage of the trial, they appear due to the transition between $2\min(P_1, P_2)$ $\max(P_1, P_2)$ when taking the minimum to find the intersection p-value.

Under the second formulation of the problem, because of the fact we can enrich into either sub-population, the form of the Oracle decision rule is quite sensitive to the value of the true treatment effects θ . The Oracle decision rule for $\theta_1 = 10$ and $\theta_2 = 8$ is shown in Figure 4-11. This Oracle decision rule is quite different to the first one we computed, there is a much larger region where the trial would enrich into the first sub-population, also since we no longer have symmetry there is a clear boundary between the first and second sub-populations.

		Formulation 1			
θ_1	θ_3	$\mathbb{P}_{\theta}(\mathcal{R}_1)$	$\mathbb{P}_{\theta}(\mathcal{R}_3)$	$\mathbb{P}_{\theta}(\mathcal{R}_b)$	$\mathbb{E}_{\theta}(G(\theta, \kappa))$
10	10	0.057	0.163	0.650	7.59
10	5	0.658	0.011	0.110	4.14
10	0	0.714	0.000	0.000	3.57

		Formulation 2			
θ_1	θ_2	$\mathbb{P}_{\theta}(\mathcal{R}_1)$	$\mathbb{P}_{\theta}(\mathcal{R}_2)$	$\mathbb{P}_{\theta}(\mathcal{R}_b)$	$\mathbb{E}_{\theta}(G(\theta, \kappa))$
10	10	0.228	0.228	0.381	6.09
10	5	0.645	0.025	0.071	3.82
10	0	0.712	0.000	0.000	3.56
5	10	0.025	0.645	0.071	3.82
0	10	0.000	0.712	0.000	3.56

Table 4.5: $\mathbb{E}_{\theta}(G(\theta, \kappa))$ of the Adaptive Enrichment design using the Oracle decision rule

Under both formulations of the problem we can simulate trials that use these Oracle decision rules to find the operating characteristics. Table 4.5 shows some examples of this using the Oracle decision rule under each formulation of the problem. Under the first formulation we see that as the treatment effect in the full population is reduced the Adaptive Enrichment trial with the Oracle decision rule places more emphasis on the sub-population of interest as this is where there is the most gain of rejecting the null hypothesis. The behaviour under the second formulation of the problem is very similar where the focus of the trial slowly moves towards a sub-population when the true treatment effect in that sub-population is higher. Part of the reason for this is that in both formulations of the problem when the treatment effect in a population drops the gain reduces, to the point that if we know the treatment effect in the full population or in the opposing sub-population is 0, there is no gain in this result and the Adaptive Enrichment trial will always choose to enrich the trial.

The Oracle decision rules are the best possible version of the Adaptive Enrichment trials giving the upper bound of maximum performance. When comparing with the fixed sampling designs we can use this to learn whether the Adaptive Enrichment trials will ever provide a benefit over the competing fixed designs. Table 4.6 shows a comparison of $\mathbb{E}_{\theta}(G(\theta, \kappa))$ between the

		Formulation 1 $\mathbb{E}_{\boldsymbol{\theta}}(G(\boldsymbol{\theta}, \kappa))$ under:		
θ_1	θ_3	FE	MH	OAE
10	10	4.50	8.59	7.59
10	5	4.50	2.82	4.14
10	0	4.50	2.48	3.57
Σ		13.50	13.89	15.30

		Formulation 2 $\mathbb{E}_{\boldsymbol{\theta}}(G(\boldsymbol{\theta}, \kappa))$ under:			
θ_1	θ_2	FE ₁	FE ₂	MH	OAE
10	10	4.50	4.50	5.89	6.09
10	5	4.50	0.91	2.84	3.82
10	0	4.50	0.00	2.62	3.56
5	10	0.91	4.50	2.84	3.82
0	10	0.00	4.50	2.62	3.56
Σ		14.41	14.41	16.81	20.85

Table 4.6: Comparing $\mathbb{E}_{\boldsymbol{\theta}}(G(\boldsymbol{\theta}, \kappa))$ for Oracle Adaptive Enrichment with fixed sampling designs: FE₁ = Fixed Enrichment in sub-population 1, FE₂ = Fixed Enrichment in sub-population 2, MH = Fixed sampling testing multiple hypotheses, OAE = Oracle Adaptive Enrichment

Adaptive Enrichment trial designs and the fixed sampling alternatives under each formulation of the problem.

Under the first formulation of the problem we see that in each of the scenarios presented the Oracle Adaptive Enrichment design is the second choice of the three trials, however the Oracle Adaptive Enrichment trial is performing more consistently over the three scenarios; this is the aim of Adaptive Enrichment, providing a compromise between the two fixed sampling methods in all cases. Taking the sum of the expected gain across this small group of possible scenarios gives a crude measure of how we perform overall. We see that while the fixed sampling trials are fairly similar to each other in terms of overall performance in this small selection of scenarios, with the Adaptive Enrichment trial providing a small overall benefit.

Making the equivalent comparisons under the second formulation of the problem. When both treatment effects are high we see that Oracle Adaptive Enrichment is the best choice of trial, this is a very encouraging result. In all other scenarios Oracle Adaptive Enrichment is the second choice and as with the first formulation of the problem this is more consistent than the fixed sampling designs. Once again taking the sum of the expected gain across all of the scenarios as a crude measure of the overall performance of the trial designs we see that Oracle Adaptive Enrichment trials perform better than any of the fixed sampling designs.

What we have seen is the best possible version of Adaptive Enrichment for each choice of θ . The results show that Adaptive Enrichment designs may provide a benefit in place of the fixed sampling designs in the examples we have considered. We know that in practice the decision at the interim analysis will not be able to be made as well and we may also wish to account for our uncertainty about the true treatment effects in our analysis of overall performance.

4.3.3 Incorporating prior uncertainty

When designing a trial it is reasonable to assume that we will not know the true treatment effect, we capture our uncertainty about the true treatment effect

as a prior distribution. Our plan is to use this prior distribution to assess trial performance and to optimise the interim decision of the Adaptive Enrichment trial under the particular choice of prior distribution and gain function. Recall from Chapter 3 that we have strong control of the FWER for all choices of $\boldsymbol{\theta}$ and so the choice of prior distribution and optimising the decision at the interim analysis based on this will not have an impact on the FWER. Our use of Bayesian methods is restricted to optimising the interim decision and evaluating the comparative performance of different designs within a framework that ensures strong control of the FWER at all times.

Let Θ be the set of all possible values of the $\boldsymbol{\theta}$ for either formulation of our problem, the cdf of the prior distribution is given by

$$\pi(\boldsymbol{\theta}) \text{ for all } \boldsymbol{\theta} \in \Theta. \quad (4.23)$$

Where before we considered the expected gain given one particular choice of the true treatment effects we may now consider the expected gain given the prior distribution $\mathbb{E}_{\pi(\boldsymbol{\theta})}(G(\boldsymbol{\theta}, \kappa))$ we will refer to this as the Bayes expected gain of the trial. The Bayes expected gain is given by

$$\mathbb{E}_{\pi(\boldsymbol{\theta})}(G(\boldsymbol{\theta}, \kappa)) = \int_{\boldsymbol{\theta} \in \Theta} \pi(\boldsymbol{\theta}) \mathbb{E}_{\boldsymbol{\theta}}(G(\boldsymbol{\theta}, \kappa)) d\boldsymbol{\theta}. \quad (4.24)$$

We seek the trial design that maximises this, as this will be the best choice for conducting a trial given our prior beliefs and the particular definition of the gain function. The same approach was recently taken by Ondra et al. (2016) for comparing the overall performance of competing trial designs in this context using their versions of utility functions, like us their aim was to compare overall trial designs to choose which is optimal.

As our plan is to use the prior distribution to help us learn which trial is most suitable this prior distribution should fully reflect our current knowledge about the true treatment effects. If we have identified sub-populations before beginning the trial it is likely we have some knowledge of what we expect the true treatment effects to be. This may be based on knowledge from earlier stages of development, knowledge of the disease being treated or based upon how the

treatment is expected to work.

For the rest of the Chapter we focus on using this framework. We shall begin by evaluating the performance of the simple Adaptive Enrichment designs used in previous examples using Equation 4.24. We shall then optimise our simple rules and assess performance increase from this optimisation. Finally we shall introduce Bayes optimal decision boundaries for the decision rule at the interim analysis, these are not restricted to any particular form.

Example

Returning again to our ongoing example, if we choose prior distributions as appropriate for each formulation of the problem we may evaluate the Bayes expected gain of our simple decision rule. We find the Bayes expected gain by Monte Carlo integration, we simulate values for $\boldsymbol{\theta}$ from the prior distribution and for each value $\boldsymbol{\theta}$ we simulate the data from a single trial, κ , to evaluate $\mathbb{I}(\mathcal{R}_1)$, $\mathbb{I}(\mathcal{R}_2)$, $\mathbb{I}(\mathcal{R}_3)$ and $\mathbb{I}(\mathcal{R}_b)$ as appropriate, from which we find the gain of that particular realisation. Averaging this across the values of $(\boldsymbol{\theta}, \kappa)$ gives our estimate of $\mathbb{E}_{\pi(\boldsymbol{\theta})}(G(\boldsymbol{\theta}, \kappa))$. In this example we simulate 1,000,000 realisations to ensure sufficient accuracy in our estimates.

In our ongoing example under the first formulation we use the prior

$$\theta_1 \sim N(9, 16), \theta_2 \sim N(3, 4).$$

where θ_1 and θ_2 are independent. This prior suggests that we believe sub-population 1 is likely to have a higher treatment effect but has a higher variance. With $\lambda = 1/2$ and $\theta_3 = \lambda\theta_1 + (1 - \lambda)\theta_2$ this is equivalent to

$$\begin{pmatrix} \theta_1 \\ \theta_3 \end{pmatrix} \sim N_2 \left(\begin{pmatrix} 9 \\ 6 \end{pmatrix}, \begin{pmatrix} 16 & 8 \\ 8 & 5 \end{pmatrix} \right).$$

Recall from Equation 4.9 that our choice of the gain function under this formulation of the problem is

$$G(\boldsymbol{\theta}, \kappa) = \lambda\theta_1\mathbb{I}(\mathcal{R}_1) + \theta_3\mathbb{I}(\mathcal{R}_3) + \theta_3\mathbb{I}(\mathcal{R}_b).$$

Formulation 1	
Trial design	$\mathbb{E}_{\pi(\boldsymbol{\theta})}(G(\boldsymbol{\theta}, \kappa))$
Fixed enrichment	3.78
Multiple hypotheses	3.74
Adaptive enrichment	3.77

Formulation 2	
Trial design	$\mathbb{E}_{\pi(\boldsymbol{\theta})}(G(\boldsymbol{\theta}, \kappa))$
Fixed enrichment	1.91
Multiple hypotheses	2.12
Adaptive enrichment	2.26

Table 4.7: Comparing $\mathbb{E}_{\pi(\boldsymbol{\theta})}(G(\boldsymbol{\theta}, \kappa))$ for fixed sampling designs with the simple adaptive enrichment design

The top half of Table 4.7 shows that under this choice of prior and gain function fixed Enrichment design is the better fixed sampling design. For the Adaptive Enrichment design our naive choice for the simple decision rule actually proves to be quite effective in this case, the Bayes expected gain of the Adaptive Enrichment design is very close to that of the fixed enrichment design.

Under the second formulation of our problem we define the independent priors for θ_1 and θ_2 as

$$\theta_1 \sim N(6, 9), \theta_2 \sim N(6, 9)$$

and use the gain function from Equation 4.10

$$G(\boldsymbol{\theta}, \kappa) = \lambda\theta_1\mathbb{I}(\mathcal{R}_1) + (1 - \lambda)\theta_2\mathbb{I}(\mathcal{R}_2) + (\lambda\theta_1 + (1 - \lambda)\theta_2)\mathbb{I}(\mathcal{R}_b).$$

The bottom half of Table 4.7 compares the Bayes expected gain for each of the fixed sampling methods and the Adaptive Enrichment trial using the simple decision rule. Due to the symmetry the fixed Enrichment design is equivalent in either sub-population, with the fixed sampling design testing multiple hypotheses being the best of the fixed sampling designs. Again our naive choice of decision rule for the Adaptive Enrichment design performs well, giving a better overall performance than both fixed sampling methods.

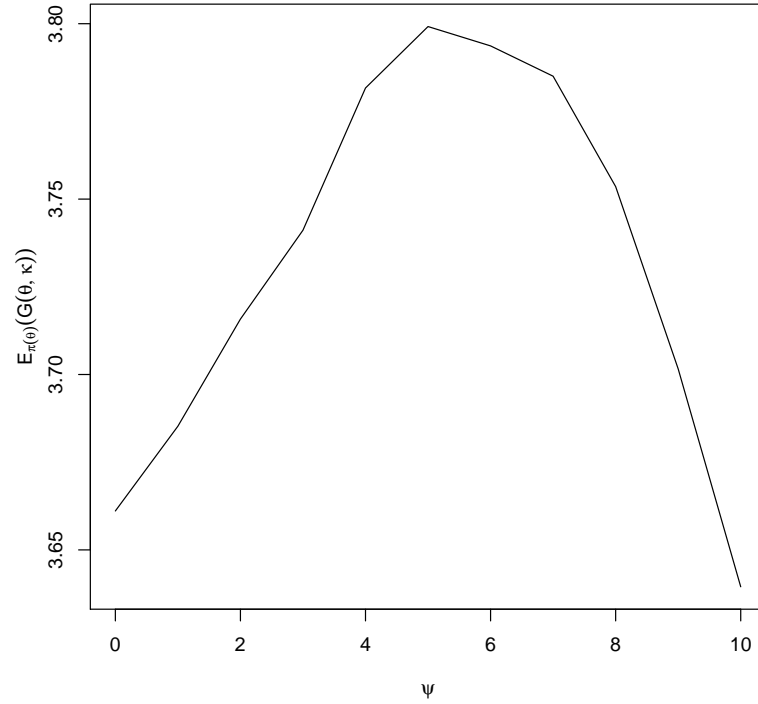


Figure 4-12: Optimising the simple decision rule under formulation 1

4.3.4 Simple decision optimisation

We can now easily optimise our simple interim decision rules. For rules of any particular form we compare the Bayes expected gain for the rule as we change the parameters that govern it and hence we may find the optimal values for these parameters.

Example

Continuing the example we will now optimise our simple decision rules for the interim analysis we have worked with up to this point in the Chapter. Recall from Equation 4.1 that in the first formulation of the problem we use a simple rule of the form

$$\text{decision} = \begin{cases} \text{continue in both populations} & \text{if } \hat{\theta}_3^{(1)} \geq \psi \\ \text{continue only in the sub-population} & \text{if } \hat{\theta}_3^{(1)} < \psi, \end{cases}$$

Formulation 1	
Trial design	$\mathbb{E}_{\pi(\boldsymbol{\theta})}(G(\boldsymbol{\theta}, \kappa))$
Fixed Enrichment	3.78
Multiple hypotheses	3.74
Adaptive Enrichment simple optimal rule	3.79
Adaptive Enrichment always enrich	3.31
Adaptive Enrichment never enrich	3.62

Formulation 2	
Trial design	$\mathbb{E}_{\pi(\boldsymbol{\theta})}(G(\boldsymbol{\theta}, \kappa))$
Fixed Enrichment	1.91
Multiple hypotheses	2.12
Adaptive Enrichment	2.29
Adaptive Enrichment always enrich	1.46
Adaptive Enrichment never enrich	2.04

Table 4.8: Comparing $\mathbb{E}_{\pi(\boldsymbol{\theta})}(G(\boldsymbol{\theta}, \kappa))$ for fixed sampling designs with the optimised simple Adaptive Enrichment design

evaluating the Bayes expected gain as we did before for $\psi = (0, 1, \dots, 10)$ we see from Figure 4-12 that this decision is optimal when $\psi = 5$. The top half of Table 4.8 shows how the Adaptive Enrichment design with the optimal simple decision rule compares to the fixed sampling designs, the Adaptive Enrichment design is now slightly better than both fixed sampling designs; in particular of the fixed sampling designs the fixed Enrichment design performs well. However the fixed Enrichment design may be a harder choice in practice (despite the statistical evidence of its performance). We will find a further improvement for the Adaptive Enrichment design when we remove the restriction of using rules of a particular form.

Similarly we can also optimise the simple rule we have been using for the second formulation of the problem, recall from equation 4.2 that we have been using a rule of the form

$$\text{decision} = \begin{cases} \text{continue in both sub-populations} & \text{if } \hat{\theta}_1^{(1)} \text{ and } \hat{\theta}_2^{(1)} \geq \psi \\ \text{continue only in sub-population 1} & \text{if } \hat{\theta}_1^{(1)} \geq \hat{\theta}_2^{(1)} \text{ and } \hat{\theta}_2^{(1)} < \psi \\ \text{continue only in sub-population 2} & \text{if } \hat{\theta}_2^{(1)} < \hat{\theta}_1^{(1)} \text{ and } \hat{\theta}_1^{(1)} < \psi. \end{cases}$$

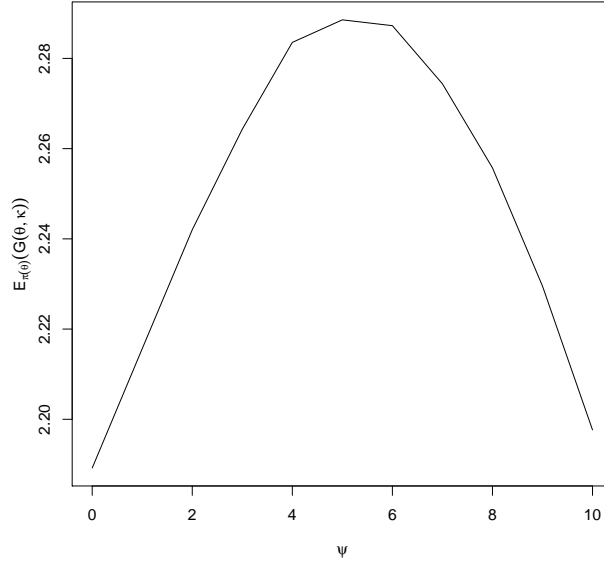


Figure 4-13: Optimising the simple decision rule under formulation 2

As before we evaluate the Bayes expected gain for $\psi = (0, 1, \dots, 10)$ giving Figure 4-13 from which we see that the simple decision is optimal when $\psi = 5$. The bottom half of Table 4.8 shows how the Adaptive Enrichment trial with the optimal simple decision rule compares to the fixed sampling designs. We see that the Adaptive Enrichment design is now slightly further ahead of the fixed sampling designs, showing an improvement of approximately 9% in the Bayes expected gain. This shows that even with a very simple decision rule the Adaptive Enrichment design is able to provide an improvement over these fixed sampling methods.

We also compare these optimised simple rules with the extremes of Adaptive Enrichment designs that always enrich the trial or never enrich the trial. From this we can clearly see that the optimised adaptive rule is providing an interesting adaptive rule, where the trial benefits from the compromise between the two extremes. We can also see that if we were in a scenario where we always made one decision the equivalent fixed sampling design would clearly be preferred but our optimised simple rule is providing a benefit for this prior.

4.4 Bayes optimal decision

Restricting the interim decision rule for the Adaptive Enrichment trial to rules of a particular form will not allow for the optimal decision to be made at all times. When constructing the Oracle decision rule, Section 4.3.1, we chose the sub-populations at the interim analysis that maximised the Bayes expected gain for the remainder of the trial given the interim observations. We apply the same method here while incorporating the prior distribution, this yields the Bayes optimal decision. Berger (2013) provides a detailed look at Bayesian decision theory in a more general setting. Schmidli et al. (2007) and Brannath et al. (2009) both seek to optimise the interim decision of adaptive trials using the Bayes predictive power. These are a special cases that are computationally simpler than the Bayes optimal rule we introduce, where the gain function is given by the probability of rejecting a null hypothesis.

For the Bayes optimal decision rule we evaluate the Bayes conditional expected gain given the observations κ_1 , at the interim analysis, $\mathbb{E}_{\pi(\boldsymbol{\theta}), \kappa_1}(G(\boldsymbol{\theta}, \kappa))$ for each of the possible interim decisions. The Bayes conditional expected gain given the observations at the interim analysis is given by

$$\mathbb{E}_{\pi(\boldsymbol{\theta}), \kappa_1}(G(\boldsymbol{\theta}, \kappa)) = \int_{\boldsymbol{\theta} \in \Theta} \pi(\boldsymbol{\theta} | \kappa_1) \mathbb{E}_{\boldsymbol{\theta}, \kappa_1}(G(\boldsymbol{\theta}, \kappa)) d\boldsymbol{\theta}. \quad (4.25)$$

The prior distribution as we have defined it relates only to the efficacy of the new treatment and thus the prior given the first stage observations may be written as

$$\begin{aligned} \pi(\boldsymbol{\theta} | \kappa_1) &= \pi(\boldsymbol{\theta} | \hat{\boldsymbol{\theta}}^{(1)}) \\ &\propto \pi(\boldsymbol{\theta}) f(\hat{\boldsymbol{\theta}}^{(1)} | \boldsymbol{\theta}), \end{aligned} \quad (4.26)$$

which is the posterior distribution given the treatment effect estimates observed in the pre-interim recruitment cohort. Having evaluated this for each of the possible decisions we choose the decision that maximises the Bayes expected gain as this is clearly the optimal decision, this defines the Bayes optimal decision rule.

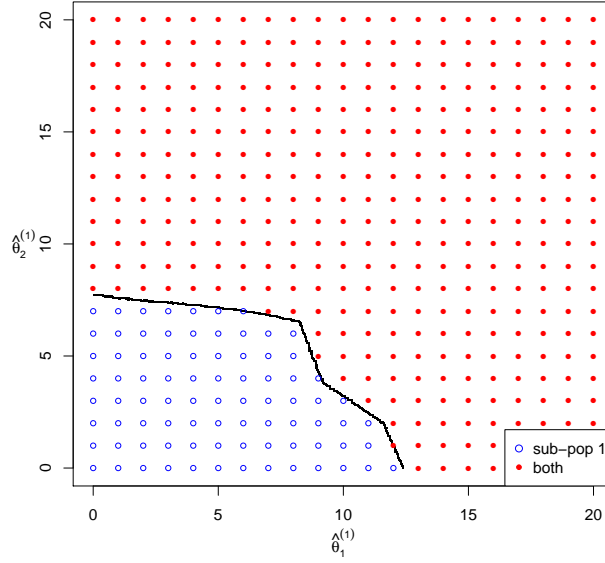


Figure 4-14: Bayes optimal decision rule under formulation 1

Example: Plotting Bayes optimal decisions

In order to evaluate the overall performance of the Bayes optimal Adaptive Enrichment designs we first need to find the optimal decision. Our overall strategy for computing the rule follows the overall approach described for the Oracle decision rule, Section 4.3.2. The difference is that for given interim observations $\hat{\boldsymbol{\theta}}^{(1)}$ we evaluate the interim Bayes expected gain given by equation 4.25. For a combination of $\hat{\boldsymbol{\theta}}^{(1)}$ under each possible decision we simulate values from the posterior, Equation 4.26, and corresponding realisations of the rest of the trial averaging to estimate the Bayes conditional expected gain for each possible decision. Even using our simplified grid scheme, Section 4.3.2, this is computationally intensive and the simulation causes the boundaries to not be plotted as smoothly as in Section 4.3.1.

Under the first formulation of the problem for our ongoing example we get the Bayes optimal decision rule shown in figure 4-14. If $\hat{\theta}_1^{(1)}$ and $\hat{\theta}_2^{(1)}$ are small then we enrich the trial, but as $\hat{\theta}_1^{(1)}$ increases we see that the trial continues in the full population for lower values of $\hat{\theta}_2^{(1)}$.

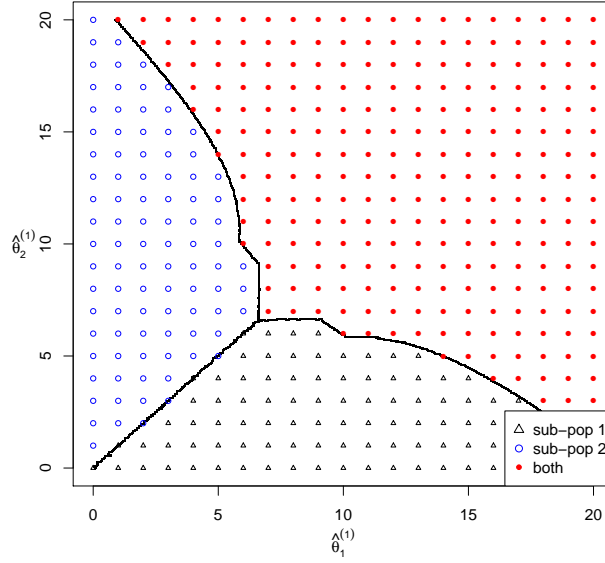


Figure 4-15: Bayes optimal decision rule under formulation 2

For the second formulation of the problem the Bayes optimal rule for our ongoing example is shown in figure 4-15. When both $\hat{\theta}_1^{(1)}$ and $\hat{\theta}_2^{(1)}$ are high we see that the trial continues recruiting from both sub-populations. For lower values of $\hat{\theta}_1^{(1)}$ and $\hat{\theta}_2^{(1)}$ the trial continues in sub-population 1 only when $\hat{\theta}_1^{(1)} > \hat{\theta}_2^{(1)}$ and continues in sub-population 2 only when $\hat{\theta}_1^{(1)} < \hat{\theta}_2^{(1)}$; this is unsurprising due to the symmetry between the sub-populations.

Example: Performance of Bayes optimal Adaptive Enrichment

Now we evaluate the Bayes expected gain for the Bayes optimal Adaptive Enrichment trials to assess the overall performance. Table 4.9 shows how the Bayes optimal Adaptive Enrichment trials compare with the fixed sampling designs for these particular examples. The Bayes expected gain of the Adaptive Enrichment trials is estimated by simulating 1,000,000 observations from the prior distributions with a corresponding realisation of the trial for each, applying an approximation of the Bayes optimal decision for each trial.

Formulation 1	
Trial design	$\mathbb{E}_{\pi(\boldsymbol{\theta})}(G(\boldsymbol{\theta}, \kappa))$
Fixed enrichment	3.78
Multiple hypotheses	3.74
Bayes optimal Adaptive enrichment	3.88

Formulation 2	
Trial design	$\mathbb{E}_{\pi(\boldsymbol{\theta})}(G(\boldsymbol{\theta}, \kappa))$
Fixed enrichment	1.91
Multiple hypotheses	2.12
Bayes optimal Adaptive enrichment	2.44

Table 4.9: Comparing $\mathbb{E}_{\pi(\boldsymbol{\theta})}(G(\boldsymbol{\theta}, \kappa))$ for fixed sampling designs with the Bayes optimal adaptive enrichment design

Under the first formulation of the problem shown in the top half of Table 4.9 we see that the Bayes optimal Adaptive Enrichment design provides a small benefit over the fixed Enrichment design which conducts the trial only using sub-population 1. This advantage of the Adaptive Enrichment design may seem slim, a large part of this is explained by the fact that whenever the trial is enriched the sample size is $3/4$ that of the fixed Enrichment design; however if for example it is not possible to achieve an agreement between all members of a trial team over which design to use, despite there being a statistical argument for doing so, presenting a compromise that on average performs slightly better than either of the fixed sampling methods in this setting is of great benefit.

Under the second formulation of the problem shown in the bottom half of Table 4.9 we see that the Bayes optimal Adaptive Enrichment design gives a large increase in overall performance compared to the fixed sampling designs. This is because the design is able to make use of beneficial characteristics of three fixed sampling designs, more regularly being able to make a useful adaptation and so the penalty in terms of overall sample size when the trial is enriched has a lower impact.

Formulation 1	
Trial design	$\mathbb{E}_{\pi(\boldsymbol{\theta})}(G(\boldsymbol{\theta}, \kappa))$
Simple Adaptive enrichment $\psi = 5$	3.80
Bayes Adaptive enrichment	3.88

Formulation 2	
Trial design	$\mathbb{E}_{\pi(\boldsymbol{\theta})}(G(\boldsymbol{\theta}, \kappa))$
Simple Adaptive enrichment $\psi = 5$	2.29
Bayes Adaptive enrichment	2.44

Table 4.10: Comparing $\mathbb{E}_{\pi(\boldsymbol{\theta})}(G(\boldsymbol{\theta}, \kappa))$ for fixed sampling designs with the Bayes optimal adaptive enrichment design

4.4.1 Comparing the Bayes optimal rule with the optimised simple decision rules

The Bayes optimal decision rule is much more computationally intensive than the optimisation of simple rules of a particular form. As such we would like to establish the benefit of this extra effort.

Example

Returning to our ongoing example we compare the Bayes optimal decision rule with the optimised simple rules we used previously. Table 4.10 shows that for our particular example Bayes optimal rule provides an improvement over the simple decision rules we proposed. However if we were to change the examples slightly we may not be certain of the best form for the simple decision rule (the ones we have used were selected with knowledge of the form of the Bayes optimal rule), the Bayes optimal decision is not limited to a particular form and requires no prior understanding of how to best make the interim decision to be applied effectively; so implementing the Bayes optimal decision rule is robust to our prior assumptions and the rest of the parameters governing our trials recruitment, always giving the best possible decision rule.

4.5 Conclusions

In this Chapter we have developed our framework for the optimisation of the decision at the interim analysis of Adaptive Enrichment trials. In Section 4.2.2 we defined the gain function that is the single measure by which we measure the performance of different designs. Using this gain function we defined the Oracle decision rule, Section 4.3.1, this allowed us to construct the best possible version of Adaptive Enrichment designs for particular choices of θ . The examples of Section 4.3.1 showed that for an appropriate range of scenarios Adaptive Enrichment designs may offer an improvement over fixed sampling designs.

To capture the true beliefs about the effectiveness of the new treatment we use prior distributions, Section 4.3.3. Using a prior distribution and a gain function we first optimise simple decision rules, however these simple decision rules are restricted to a particular form and so not optimal in all cases. In Section 4.4 we construct Bayes optimal Adaptive Enrichment designs which use the best possible decision for the combination of prior distribution and gain function. Using Bayes optimal Adaptive Enrichment designs we find that for our particular examples the Adaptive Enrichment designs have a higher overall performance than the fixed sampling designs we compare them with. Under the first formulation of the problem the fixed Enrichment design still performs quite well, but the option of only conducting the trial in one sub-population will have less appeal in practice in addition to the observed lower overall performance.

Chapter 5

Delayed response and survival endpoints

5.1 Adaptive Enrichment with delayed response

5.1.1 Recruitment

Up to this point we have assumed that the data from the patients in the pre-interim recruitment cohort are fully available whenever we wish to conduct the interim analysis in an Adaptive Enrichment trial, however this will rarely be the case. We assumed that the time between administering a treatment and observing the response was negligible, but in practice we may wait weeks or months before the response is planned to be observed. For the fixed sampling designs there is no change as the analysis of the trial is conducted after all data collection has been completed, however there are some changes to the Adaptive Enrichment trials.

Suppose we recruit n patients recall that for $\tau \in [0, 1]$, $n^{(1)} = \tau n$ patients are recruited before the interim analysis, the pre-interim recruitment cohort, and $n^{(2)} = (1 - \tau)n$ patients are recruited after the interim analysis, the post-interim recruitment cohort. At the time of the interim analysis for $\tilde{\tau} \in [0, \tau]$ only $\tilde{n}^{(1)} = \tilde{\tau}n$ patients have had their response to the treatment observed. For the final analysis and hypothesis testing we have the observations for the remaining $n^{(1)} - \tilde{n}^{(1)}$ patients from the pre-interim recruitment cohort. There is no impact on the post-interim recruitment cohort we will still be able to make a choice over

how this is recruited, as previously, and we assume all observations from both recruitment cohorts are available at the final analysis.

Under both formulations of the problem at the interim analysis we have treatment effect estimates $\hat{\theta}_1^{(\iota)}$ in the first sub-population and $\hat{\theta}_2^{(\iota)}$ in the second sub-population from the patients in the pre-interim recruitment cohort who we have fully observed by the interim analysis. Under the first formulation we have $\hat{\theta}_3^{(\iota)}$ where as usual $\hat{\theta}_3^{(\iota)} = \lambda\hat{\theta}_1^{(\iota)} + (1 - \lambda)\hat{\theta}_2^{(\iota)}$. Using the same $\tilde{\mathcal{I}}$ as previously these estimates follow known distributions,

$$\hat{\theta}_1^{(\iota)} \sim N(\theta_1, (\lambda\tilde{\tau}\tilde{\mathcal{I}})^{-1}) \text{ and } \hat{\theta}_2^{(\iota)} \sim N(\theta_2, ((1 - \lambda)\tilde{\tau}\tilde{\mathcal{I}})^{-1}). \quad (5.1)$$

$\hat{\theta}_1^{(\iota)}$ and $\hat{\theta}_3^{(\iota)}$ have the familiar joint distribution given by

$$\begin{pmatrix} \hat{\theta}_1^{(\iota)} \\ \hat{\theta}_3^{(\iota)} \end{pmatrix} \sim N_2 \left(\begin{pmatrix} \theta_1 \\ \theta_3 \end{pmatrix}, \begin{pmatrix} (\lambda\tilde{\tau}\tilde{\mathcal{I}})^{-1} & (\tilde{\tau}\tilde{\mathcal{I}})^{-1} \\ (\tilde{\tau}\tilde{\mathcal{I}})^{-1} & (\tilde{\tau}\tilde{\mathcal{I}})^{-1} \end{pmatrix} \right).$$

For the remainder of the trial we continue to collect observations from the pre-interim recruitment cohort. We estimate the treatment effect from the observations in the pre-interim recruitment cohort who are observed after the interim analysis. These treatment effects are $\hat{\theta}_1^{(\neg\iota)}$ in the first sub-population and $\hat{\theta}_2^{(\neg\iota)}$ in the second sub-population under both formulations with $\hat{\theta}_3^{(\neg\iota)} = \lambda\hat{\theta}_1^{(\neg\iota)} + (1 - \lambda)\hat{\theta}_2^{(\neg\iota)}$ the estimate of the treatment effect in the full population for the first formulation of the problem. The distributions are

$$\hat{\theta}_1^{(\neg\iota)} \sim N(\theta_1, (\lambda(\tau - \tilde{\tau})\tilde{\mathcal{I}})^{-1}), \quad \hat{\theta}_2^{(\neg\iota)} \sim N(\theta_2, ((1 - \lambda)(\tau - \tilde{\tau})\tilde{\mathcal{I}})^{-1}). \quad (5.2)$$

where $\hat{\theta}_1^{(\neg\iota)}$ and $\hat{\theta}_2^{(\neg\iota)}$ are independent and

$$\begin{pmatrix} \hat{\theta}_1^{(\neg\iota)} \\ \hat{\theta}_3^{(\neg\iota)} \end{pmatrix} \sim N_2 \left(\begin{pmatrix} \theta_1 \\ \theta_3 \end{pmatrix}, \begin{pmatrix} (\lambda(\tau - \tilde{\tau})\tilde{\mathcal{I}})^{-1} & ((\tau - \tilde{\tau})\tilde{\mathcal{I}})^{-1} \\ ((\tau - \tilde{\tau})\tilde{\mathcal{I}})^{-1} & ((\tau - \tilde{\tau})\tilde{\mathcal{I}})^{-1} \end{pmatrix} \right).$$

For all populations, the treatment effect estimate from the first stage of the trial for the final analysis is

$$\hat{\theta}_i^{(1)} = \frac{\tilde{\tau}}{\tau}\hat{\theta}_i^{(\iota)} + \frac{\tau - \tilde{\tau}}{\tau}\hat{\theta}_i^{(\neg\iota)} \text{ for } i = 1, 2 \text{ or } 3. \quad (5.3)$$

The estimates of the treatment effects in the post-interim recruitment cohort are unchanged and follow the distributions given in Section 3.6.

5.1.2 Hypothesis testing

The hypothesis testing may be conducted in the same way as we defined in Section 3.5.2. We use the estimates of the treatment effects from the pre-interim recruitment cohort $\hat{\boldsymbol{\theta}}^{(1)}$ and the post-interim recruitment cohort $\hat{\boldsymbol{\theta}}^{(2)}$ to construct these hypotheses tests. These treatment effects have the same distributions as stated previously in Section 3.6. Recall that for the pre-interim recruitment cohort we have

$$\hat{\theta}_1^{(1)} \sim N(\theta_1, (\lambda\tau\tilde{\mathcal{I}})^{-1})$$

for the sub-population 1,

$$\hat{\theta}_2^{(1)} \sim N(\theta_2, ((1-\lambda)\tau\tilde{\mathcal{I}})^{-1})$$

for sub-population 2, under the first formulation we also have

$$\begin{pmatrix} \hat{\theta}_1^{(1)} \\ \hat{\theta}_3^{(1)} \end{pmatrix} \sim N_2 \left(\begin{pmatrix} \theta_1 \\ \theta_3 \end{pmatrix}, \begin{pmatrix} (\lambda\tau\tilde{\mathcal{I}})^{-1} & (\tau\tilde{\mathcal{I}})^{-1} \\ (\tau\tilde{\mathcal{I}})^{-1} & (\tau\tilde{\mathcal{I}})^{-1} \end{pmatrix} \right).$$

The distributions of the treatment effects for the post-interim recruitment cohort depend on which decision has been made, if we continue in both sub-populations we get

$$\hat{\theta}_1^{(2)} \sim N(\theta_1, (\lambda(1-\tau)\tilde{\mathcal{I}})^{-1})$$

for sub-population 1,

$$\hat{\theta}_2^{(2)} \sim N(\theta_2, ((1-\lambda)(1-\tau)\tilde{\mathcal{I}})^{-1})$$

for sub-population 2, under the first formulation we also have

$$\begin{pmatrix} \hat{\theta}_1^{(2)} \\ \hat{\theta}_3^{(2)} \end{pmatrix} \sim N_2 \left(\begin{pmatrix} \theta_1 \\ \theta_3 \end{pmatrix}, \begin{pmatrix} (\lambda(1-\tau)\tilde{\mathcal{I}})^{-1} & ((1-\tau)\tilde{\mathcal{I}})^{-1} \\ ((1-\tau)\tilde{\mathcal{I}})^{-1} & ((1-\tau)\tilde{\mathcal{I}})^{-1} \end{pmatrix} \right).$$

if we continue only in sub-population $i = 1, 2$ we have

$$\hat{\theta}_i^{(2)} \sim N(\theta_i, ((1 - \tau)\tilde{\mathcal{I}})^{-1})$$

As usual we require strong control of the FWER as which we defined in Equation 3.1 to be

$$\mathbb{P}_{\boldsymbol{\theta}}(\text{Reject at least one true null hypothesis}) \leq \alpha \text{ for all } \boldsymbol{\theta}.$$

but the estimates of the treatment effects at the final analysis follow the same distributions and so we may use the same hypothesis testing methods described in Section 3.5.2. In our examples we use a closed testing procedure, Section 3.2.2, finding intersection P-values within each recruitment cohort using Simes method from Equation 3.4, then using a combination test such as the weighted inverse normal, Section 3.4.1, to find combined p-values across the whole trial which define which hypotheses may be rejected globally.

5.1.3 Optimisation

Optimisation of the Adaptive Enrichment design changes slightly with delayed response. We still want to make the decision that optimises the Bayes expected gain given what we have observed at the interim analysis, in Chapter 4 we used equation 4.25,

$$\mathbb{E}_{\pi(\boldsymbol{\theta}), \kappa_1}(G(\boldsymbol{\theta}, \kappa)) = \int_{\boldsymbol{\theta} \in \Theta} \pi(\boldsymbol{\theta} | \kappa_1) \mathbb{E}_{\boldsymbol{\theta}, \kappa_1}(G(\boldsymbol{\theta}, \kappa)) d\boldsymbol{\theta},$$

to find the expected gain of each action at the interim analysis and chose the decision that maximised this. With a delayed response we make a small change to our notation. Under immediate response we evaluated this given all observations from the pre-interim recruitment cohort κ_1 , now we may only use the observations we have available from the pre-interim recruitment cohort at the time of interim analysis, κ_ℓ say. We evaluate the Bayes conditional expected gain given the available observations $\mathbb{E}_{\pi(\boldsymbol{\theta}), \kappa_\ell}(G(\boldsymbol{\theta}, \kappa))$, this is

$$\mathbb{E}_{\pi(\boldsymbol{\theta}), \kappa_\ell}(G(\boldsymbol{\theta}, \kappa)) = \int_{\boldsymbol{\theta} \in \Theta} \pi(\boldsymbol{\theta} | \kappa_\ell) \mathbb{E}_{\boldsymbol{\theta}, \kappa_\ell}(G(\boldsymbol{\theta}, \kappa)) d\boldsymbol{\theta}. \quad (5.4)$$

We may replace $\pi(\boldsymbol{\theta}|\kappa_\iota)$ by a posterior distribution given the estimates of the treatment effect from the observed data $\hat{\boldsymbol{\theta}}^{(\iota)}$,

$$\begin{aligned}\pi(\boldsymbol{\theta}|\kappa_\iota) &= \pi(\boldsymbol{\theta}|\hat{\boldsymbol{\theta}}^{(\iota)}) \\ &\propto \pi(\boldsymbol{\theta})f(\hat{\boldsymbol{\theta}}^{(\iota)}|\boldsymbol{\theta}),\end{aligned}$$

Additionally $\mathbb{E}_{\boldsymbol{\theta}|\kappa_\iota}(G(\boldsymbol{\theta}, \kappa))$ now depends not only on the treatment effect estimates from the post-interim recruitment cohort, $\hat{\boldsymbol{\theta}}^{(2)}$, but also on the treatment effect estimate from the patients in the pre-interim recruitment cohort who have not been observed at the interim analysis $\hat{\boldsymbol{\theta}}^{(\iota)}$. So when simulating the possible outcomes for the remainder of the trial to evaluate the expected performance of each possible decision we now simulate values for $\hat{\boldsymbol{\theta}}^{(\iota)}$ in addition to $\hat{\boldsymbol{\theta}}^{(2)}$.

5.1.4 Example

As we did in Chapter 4 we will continue to return to the same core example as we introduce new ideas. To maintain consistency we use the same choices as before. These were: a fixed value of $\tilde{\mathcal{I}}$,

$$\tilde{\mathcal{I}} = \left(\frac{\Phi^{-1}(0.9) + \Phi^{-1}(0.975)}{10} \right)^2$$

chosen such that a test based on $\hat{\theta} \sim N(\theta, \mathcal{I}^{-1})$ has power of 0.9 when $\theta = 10$ for testing a single null hypothesis, say $H_0 : \theta \leq 0$ at $\alpha = 0.025$; for the designs that make use of sub-populations we set $\lambda = 0.5$, meaning that the first sub-population accounts for half of the total sample; the Adaptive Enrichment designs will also have multiple stages, we set $\tau = 0.5$ giving half of the sample in each stage of the trial. With these defined we may use the distributions given in Section 5.1 for each of the possible trial designs. We conduct the hypothesis tests using the same methods as described in Section 3.6 setting $\alpha = 0.025$.

We will make use of the same gain functions and priors in our optimisation framework as we did in the examples of Chapter 4. For the first formulation of

the problem we had the gain function given in Equation 4.5,

$$G(\boldsymbol{\theta}, \kappa) = \lambda\theta_1\mathbb{I}(\mathcal{R}_1) + \theta_3\mathbb{I}(\mathcal{R}_3) + \theta_3\mathbb{I}(\mathcal{R}_b),$$

and the prior distributions were

$$\theta_1 \sim N(9, 16) \text{ and } \theta_2 \sim N(3, 4)$$

where θ_1 and θ_2 are independent. Since $\lambda = 1/2$ we have $\theta_3 = \lambda\theta_1 + (1 - \lambda)\theta_2$ and may write the joint distribution as

$$\begin{pmatrix} \theta_1 \\ \theta_3 \end{pmatrix} \sim N_2 \left(\begin{pmatrix} 9 \\ 6 \end{pmatrix}, \begin{pmatrix} 16 & 8 \\ 8 & 5 \end{pmatrix} \right).$$

Under the second formulation of the problem we define the gain function as in equation 4.6,

$$G(\boldsymbol{\theta}, \kappa) = \lambda\theta_1\mathbb{I}(\mathcal{R}_1) + (1 - \lambda)\theta_2\mathbb{I}(\mathcal{R}_2) + (\lambda\theta_1 + (1 - \lambda)\theta_2)\mathbb{I}(\mathcal{R}_b).$$

For the second formulation of the problem we chose symmetric independent prior distributions

$$\theta_1 \sim N(6, 9) \text{ and } \theta_2 \sim N(6, 9).$$

Plotting the Bayes optimal decision boundaries

To plot the Bayes optimal decision rule we again use the same sub dividing strategy as we described for the Oracle decision rule, Section 4.3.2. For each grid point where we use simulation. We simulate 500,000 realisations of $\boldsymbol{\theta}$, then for each possible decision we simulate 500,000 realisations of $\hat{\boldsymbol{\theta}}^{(-i)}$ and $\hat{\boldsymbol{\theta}}^{(2)}$ estimating $\mathbb{E}_{\boldsymbol{\theta}, \kappa_i}(G(\boldsymbol{\theta}, \kappa))$ for each possible decision. The grid point is assigned the decision that gives the highest estimate of $\mathbb{E}_{\boldsymbol{\theta}, \kappa_i}(G(\boldsymbol{\theta}, \kappa))$.

Setting $\tilde{\tau} = 0.3$ we illustrate the Bayes optimal decision rule under the first formulation in Figure 5-1. We see that the Bayes optimal boundary has shifted when compared to the equivalent boundary for immediate response that we saw in Figure 4-14, under this delayed response the estimates of treatment effects

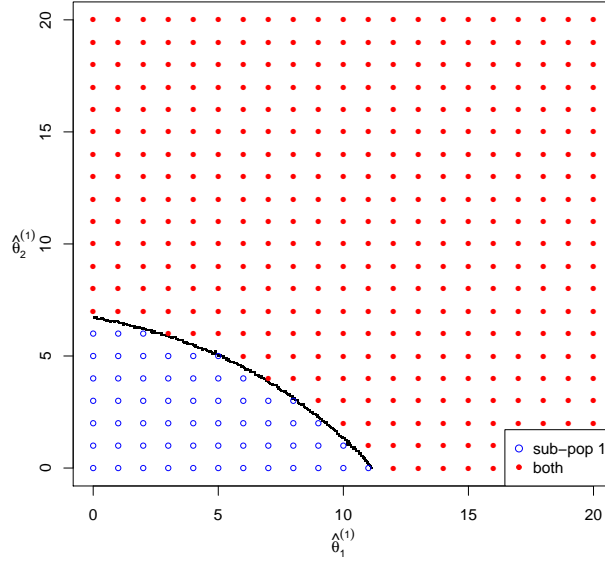


Figure 5-1: Bayes optimal decision boundary for a delayed response under formulation 1

must be lower in order for the trial to be enriched; this decision rule is more similar to fixed sampling design testing multiple hypotheses, choosing to enrich the trial less often.

With the same value of $\tilde{\tau} = 0.3$ the Bayes optimal decision boundary for the second formulation of the problem is shown in Figure 5-2. Again the trial continues in both sub-populations for lower values of the treatment effect estimates. Under both formulations of the problem this change in the decision rule reflects the fact that under the prior distribution this is the favoured decision and we have less information from the trial to base our decision on due to the delayed response.

Overall trial performance

To estimate $\mathbb{E}_{\boldsymbol{\theta}}(G(\boldsymbol{\theta}, \kappa))$ in the setting of delayed response we start by simulating 1,000,000 realisations of $\boldsymbol{\theta}$ and $\hat{\boldsymbol{\theta}}^{(i)}$, then applying an approximation of the Bayes optimal decision we simulate the 1,000,000 realisations of $\hat{\boldsymbol{\theta}}^{(-i)}$ and $\hat{\boldsymbol{\theta}}^{(2)}$ to find the outcome for 1,000,000 realisations of the trial, averaging the gain

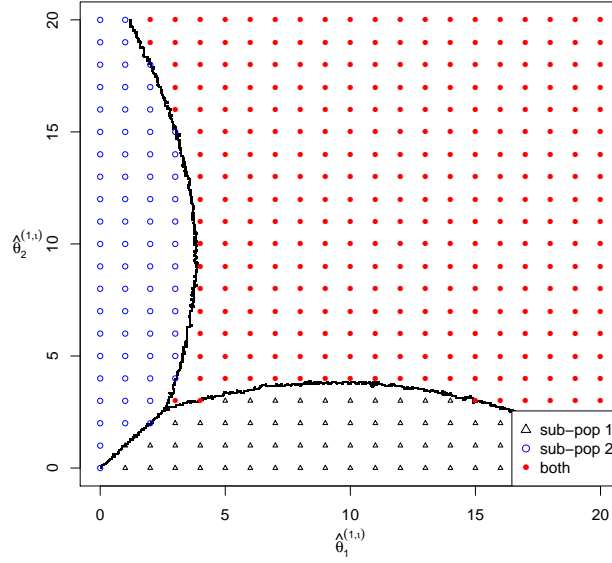


Figure 5-2: Bayes optimal decision boundary for a delayed response under formulation 2

across all realisations gives us our estimate of $\mathbb{E}_{\boldsymbol{\theta}, \kappa_L}(G(\boldsymbol{\theta}, \kappa))$.

Even though we are still able to optimise the Adaptive Enrichment designs when using a delayed response the important question is how has the delayed response affected the properties of the Adaptive Enrichment design? Table 5.1 compares the Bayes expected gain of the Bayes optimal Adaptive Enrichment design with the fixed sampling designs as we vary the delay in response by changing $\tilde{\tau}$.

Under the first formulation of the problem we see in Table 5.1 that as we decrease $\tilde{\tau}$ the expected gain and the probability of enriching the trial in the Bayes optimal Adaptive Enrichment design decrease. For all of the value of $\tilde{\tau} < 0.5$ that we have simulated we see that the fixed Enrichment design gives a higher Bayes expected gain and for $\tilde{\tau} = 0.1$ the expected gain of the Bayes optimal Adaptive Enrichment trial is the lowest of all three options. This shows that the impact of the loss of information at the interim analysis of the Adaptive Enrichment design can severely impact on the ability to make good interim decisions about a trials

Formulation 1		
Trial type	% enriched	$\mathbb{E}_{\pi(\boldsymbol{\theta})}(G(\boldsymbol{\theta}, \kappa))$
Adaptive Enrichment:		
$\tilde{\tau} = 0.1$	20.9%	3.66
$\tilde{\tau} = 0.2$	33.2%	3.69
$\tilde{\tau} = 0.3$	40.1%	3.73
$\tilde{\tau} = 0.4$	44.7%	3.77
$\tilde{\tau} = 0.5$	45.0%	3.88
Fixed Samples:		
Fixed Enrichment	-	3.78
Multiple hypotheses	-	3.67
Formulation 2		
Trial type	% enriched	$\mathbb{E}_{\pi(\boldsymbol{\theta})}(G(\boldsymbol{\theta}, \kappa))$
Adaptive Enrichment:		
$\tilde{\tau} = 0.1$	22.3%	2.07
$\tilde{\tau} = 0.2$	46.0%	2.13
$\tilde{\tau} = 0.3$	59.4%	2.19
$\tilde{\tau} = 0.4$	67.4%	2.25
$\tilde{\tau} = 0.5$	70.8%	2.44
Fixed Samples:		
Fixed Enrichment	-	1.91
Multiple hypotheses	-	2.12

Table 5.1: Comparing $\mathbb{E}_{\pi(\boldsymbol{\theta})}(G(\boldsymbol{\theta}, \kappa))$ for Bayes optimal Adaptive Enrichment designs with a varying proportion of interim observations

recruitment.

Under the second formulation of the problem we see in Table 5.1 that the Bayes optimal Adaptive Enrichment for this particular set of examples performs well. The Bayes optimal Adaptive Enrichment trial is only worse than one of the fixed sampling alternatives (testing multiple hypotheses) when $\tilde{\tau} = 0.1$. As discussed previously this is because the Adaptive Enrichment design is able to pick from three decisions where it offers good performance.

5.2 Longitudinal analysis

5.2.1 Data collection

When conducting a trial it is routine to collect observations about other variables in addition to the primary endpoint for the final analysis. Many of these additional variables will be things such as stratification factors or safety endpoints but it is also commonplace to collect data on endpoints similar to the primary endpoint. Suppose, for example, that the primary endpoint is measured after 6 months, during the trial the same measurement is taken after only a single month, a secondary endpoint. If the interim analysis were conducted after a year those recruited within the first 6 months would have had both the short and long term endpoints observed, however there are more data available from the short term observations of the patients recruited between 6 and 11 months. Assuming we are unable to use the short term endpoint in the final analysis (the long term endpoint may be a regulatory requirement), we may still use these extra observations to enhance our interim decision making. This will not impact the FWER; as discussed in Section 3.5.2 we achieve strong control of the FWER no matter how the interim decision is made. Recall that this is achieved by using an overall closed testing procedure that defines the possible choices at the interim analysis before the trial begins.

Let the treatment effects we have used so far in each sub-population, θ_1 and θ_2 , be the long term treatment effects, measured by the primary endpoint. We define ν_1 to be the short term treatment effect in the first sub-population and

ν_2 to be the short term treatment effect in the second sub-population. Suppose at the interim analysis we have recruited $n^{(1)}$ patients, let n_ν be the number of patients who have observations for the short term endpoint then for τ_ν such that

$$\tau_\nu n_\nu = \frac{\tilde{\tau}}{\tau} n^{(1)}.$$

At the interim analysis we have $\tau_\nu n_\nu$ patients who have observations of both the primary endpoint and the secondary endpoint giving estimates $\hat{\theta}_1^{(\iota)}$ and $\hat{\theta}_2^{(\iota)}$ for the long term treatment effect and estimates $\hat{\nu}_1^{(1)}$ and $\hat{\nu}_2^{(1)}$ for the short term treatment effect respectively; we assume that the primary and secondary endpoints are correlated within each patient, taking the same correlation for both sub-populations $\hat{\theta}_i^{(\iota)}$ and $\hat{\nu}_i^{(1)}$ have correlation ρ for $i = 1, 2$. The other $(1 - \tau_\nu)n_\nu$ patients who have observations for the secondary endpoint at the interim analysis give estimates for the short term treatment effect $\hat{\nu}_1^{(2)}$ and $\hat{\nu}_2^{(2)}$.

As we did in Section 3.6 we assume a common variance for the short term endpoint for the patients observations in both sub-populations and both treatment groups, σ_ν^2 say, and define

$$\mathcal{I}_\nu = \frac{n_\nu}{4\sigma_\nu^2}.$$

With λ as the proportion of patients in sub-population 1 as usual, the marginal distributions for the estimates of ν_1 and ν_2 are: for the patients with both long term and short term observations available at the interim analysis

$$\hat{\nu}_1^{(1)} \sim N(\nu_1, (\tau_\nu \lambda \mathcal{I}_\nu)^{-1}), \hat{\nu}_2^{(1)} \sim N(\nu_2, (\tau_\nu (1 - \lambda) \mathcal{I}_\nu)^{-1})$$

and for the patients with only the short term observations at the interim analysis

$$\hat{\nu}_1^{(2)} \sim N(\nu_1, ((1 - \tau_\nu) \lambda \mathcal{I}_\nu)^{-1}), \hat{\nu}_2^{(2)} \sim N(\nu_2, ((1 - \tau_\nu) (1 - \lambda) \mathcal{I}_\nu)^{-1}).$$

From Equation 5.1 we know that the estimates of the long term treatment effect for the patients who have observations of the primary endpoint at the interim

analysis have the marginal distributions

$$\hat{\theta}_1^{(\iota)} \sim N(\theta_1, (\lambda\tilde{\tau}\tilde{\mathcal{I}})^{-1}), \hat{\theta}_2^{(\iota)} \sim N(\theta_2, ((1-\lambda)\tilde{\tau}\tilde{\mathcal{I}})^{-1}),$$

where $\hat{\theta}_1^{(\iota)}$ and $\hat{\theta}_2^{(\iota)}$ are independent. In addition we now have a correlation between the short term and long term estimates,

$$\text{corr}(\hat{\theta}_1^{(\iota)}, \hat{\nu}_1^{(1)}) = \text{corr}(\hat{\theta}_2^{(\iota)}, \hat{\nu}_2^{(1)}) = \rho.$$

For the patients in the pre-interim recruitment cohort who do not contribute a long term observation at the interim analysis we continue observing the primary endpoint. From these patients there are $(1-\tau_\nu)n_\nu$ who had an observation of the secondary endpoint at the interim analysis from which we find $\hat{\theta}_1^{(\neg\iota,1)}$ and $\hat{\theta}_2^{(\neg\iota,1)}$; while the remaining $n^{(1)} - n_\nu$ patients did not giving $\hat{\theta}_1^{(\neg\iota,2)}$ and $\hat{\theta}_2^{(\neg\iota,2)}$. Let

$$\zeta = \frac{\tilde{\tau}(1-\tau_\nu)}{\tau_\nu(\tau-\tilde{\tau})},$$

note that we have $0 \leq \zeta \leq 1$ since

$$\tau_\nu \geq \frac{\tilde{\tau}}{\tau}$$

by definition. The marginal distributions of $\hat{\theta}_1^{(\neg\iota,1)}$ and $\hat{\theta}_2^{(\neg\iota,1)}$ are

$$\hat{\theta}_1^{(\neg\iota,1)} \sim N(\theta_1, (\zeta(\tau-\tilde{\tau})\lambda\mathcal{I})^{-1}), \hat{\theta}_2^{(\neg\iota,1)} \sim N(\theta_2, (\zeta(\tau-\tilde{\tau})\lambda\mathcal{I})^{-1}),$$

where as previously

$$\text{corr}(\hat{\theta}_1^{(\iota,1)}, \hat{\nu}_1^{(2)}) = \text{corr}(\hat{\theta}_2^{(\iota,1)}, \hat{\nu}_2^{(2)}) = \rho.$$

The marginal distributions for $\hat{\theta}_1^{(\neg\iota,2)}$ and $\hat{\theta}_2^{(\neg\iota,2)}$ are

$$\hat{\theta}_1^{(\neg\iota,2)} \sim N(\theta_1, ((1-\zeta)(\tau-\tilde{\tau})\lambda\mathcal{I})^{-1}), \hat{\theta}_2^{(\neg\iota,2)} \sim N(\theta_2, ((1-\zeta)(\tau-\tilde{\tau})(1-\lambda)\mathcal{I})^{-1}). \quad (5.5)$$

We have that $\hat{\theta}_i^{(\neg\iota)} = \zeta\hat{\theta}_i^{(\neg\iota,1)} + (1-\zeta)\hat{\theta}_i^{(\neg\iota,2)}$ for $i = 1, 2$, where $\hat{\theta}_1^{(\neg\iota)}$ and $\hat{\theta}_2^{(\neg\iota)}$ follow the distributions from Equation 5.2.

5.2.2 Interim optimisation

At the time of the interim analysis under both formulations of the problem we will have observed $\hat{\nu}_1^{(1)}$, $\hat{\nu}_2^{(1)}$, $\hat{\theta}_1^{(\iota)}$, $\hat{\theta}_2^{(\iota)}$, $\hat{\nu}_1^{(2)}$ and $\hat{\nu}_1^{(2)}$. For either the first or second sub-population we can enhance the interim estimate of θ_i for $i = 1$ or 2 respectively by following the method described by Hampson and Jennison (2013). We may write the joint distribution within a single sub-population as

$$\begin{pmatrix} \hat{\nu}_i^{(1)} \\ \hat{\theta}_i^{(\iota)} \\ \hat{\nu}_i^{(2)} \end{pmatrix} \sim N_3 \left(\begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} \nu_i \\ \theta_i \end{pmatrix}, \begin{pmatrix} v_{11} & v_{12} & 0 \\ v_{12} & v_{22} & 0 \\ 0 & 0 & v_{33} \end{pmatrix} \right) \quad (5.6)$$

Notice that if write

$$Y = \begin{pmatrix} \hat{\nu}_i^{(1)} \\ \hat{\theta}_i^{(\iota)} \\ \hat{\nu}_i^{(2)} \end{pmatrix}, X = \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 0 \end{pmatrix}, \boldsymbol{\theta} = \begin{pmatrix} \nu_i \\ \theta_i \end{pmatrix}, \Sigma = \begin{pmatrix} v_{11} & v_{12} & 0 \\ v_{12} & v_{22} & 0 \\ 0 & 0 & v_{33} \end{pmatrix}$$

then equation 5.6 can be written as

$$Y \sim N(X\boldsymbol{\theta}, \Sigma),$$

which is a normal linear model which we may fit by maximum likelihood

$$\hat{\boldsymbol{\theta}} = (X^T \Sigma^{-1} X)^{-1} X^T \Sigma^{-1} Y. \quad (5.7)$$

This will give us the vector of parameter estimates

$$\hat{\boldsymbol{\theta}} = \begin{pmatrix} \hat{\theta}'_i \\ \hat{\nu}_i \end{pmatrix}$$

where $\hat{\theta}'_i$ is an improved estimate of θ_i at the interim analysis.

Noting that $v_{12} = \rho\sqrt{v_{11}v_{22}}$ and solving equation 5.7 we find that

$$\hat{\theta}'_i = \hat{\theta}_i^{(2)} + (\hat{\nu}^{(2)} - \hat{\nu}^{(1)}) \frac{\rho\sqrt{v_{11}v_{22}}}{v_{11} + v_{33}}$$

and

$$\hat{\nu}_i = \frac{v_{33}\hat{\nu}^{(1)} + v_{11}\hat{\nu}^{(2)}}{v_{11} + v_{33}}.$$

The joint distribution of these estimates is given by

$$\begin{pmatrix} \hat{\theta}'_i \\ \hat{\nu}_i \end{pmatrix} \sim N_2 \left(\begin{pmatrix} \theta \\ \nu \end{pmatrix}, \frac{1}{v_{11} + v_{33}} \begin{pmatrix} v_{22}(v_{11} + v_{33}) - \rho^2 v_{11} v_{22} & v_{33} \rho \sqrt{v_{11} v_{22}} \\ v_{33} \rho \sqrt{v_{11} v_{22}} & v_{11} v_{33} \end{pmatrix} \right) \quad (5.8)$$

For the distributions we defined in Section 5.2.1 we have: in sub-population 1

$$v_{11} = (\tau_\nu \lambda \mathcal{I}_\nu)^{-1}, \quad v_{22} = (\zeta(\tau - \tilde{\tau}) \lambda \mathcal{I})^{-1}, \quad v_{33} = ((1 - \tau_\nu) \lambda \mathcal{I}_\nu)^{-1}$$

and in sub-population 2

$$v_{11} = (\tau_\nu (1 - \lambda) \mathcal{I}_\nu)^{-1}, \quad v_{22} = (\zeta(\tau - \tilde{\tau}) (1 - \lambda) \mathcal{I})^{-1}, \quad v_{33} = ((1 - \tau_\nu) (1 - \lambda) \mathcal{I}_\nu)^{-1}$$

We may now use our usual method for optimisation, evaluating the expected gain for each possible decision and choosing the decision that maximises this. To evaluate the expected gain at the interim analysis we use Equation 5.4,

$$\mathbb{E}_{\pi(\boldsymbol{\theta}), \kappa_\iota} (G(\boldsymbol{\theta}, \kappa)) = \int_{\boldsymbol{\theta} \in \Theta} \pi(\boldsymbol{\theta} | \kappa_\iota) \mathbb{E}_{\boldsymbol{\theta}, \kappa_\iota} (G(\boldsymbol{\theta}, \kappa)) d\boldsymbol{\theta}.$$

The prior distribution $\pi(\boldsymbol{\theta})$ should be replaced by a prior on both $\boldsymbol{\theta}$ and $\boldsymbol{\nu}$, say $\pi(\boldsymbol{\theta}, \boldsymbol{\nu})$. This prior is updated based on the estimate of $\boldsymbol{\nu}$ and the improved estimate of $\boldsymbol{\theta}$, hence the posterior $\pi(\boldsymbol{\theta} | \kappa)$ is given by $\pi(\boldsymbol{\theta}, \boldsymbol{\nu} | \hat{\boldsymbol{\theta}}', \hat{\boldsymbol{\nu}})$.

As with the optimisation of delayed response from Section 5.1.3, $\mathbb{E}_{\boldsymbol{\theta}, \kappa_\iota} (G(\boldsymbol{\theta}, \kappa))$ will depend on the distributions of the values yet to be observed $\hat{\theta}_i^{(\neg \iota)}$ and $\hat{\theta}_i^{(2)}$ for $i = 1, 2$. Recall that $\hat{\theta}_i^{(\neg \iota)} = \zeta \hat{\theta}_i^{(\neg \iota, 1)} + (1 - \zeta) \hat{\theta}_i^{(\neg \iota, 2)}$ for $i = 1, 2$, where

$$\hat{\theta}_1^{(\neg \iota, 2)} \sim N(\theta_1, ((1 - \zeta)(\tau - \tilde{\tau}) \lambda \mathcal{I})^{-1}), \quad \hat{\theta}_2^{(\neg \iota, 2)} \sim N(\theta_2, ((1 - \zeta)(\tau - \tilde{\tau}) (1 - \lambda) \mathcal{I})^{-1}).$$

All that is left is the conditional distributions of $\hat{\theta}_i^{(\neg \iota, 1)} | \hat{\nu}^{(2)}$ for $i = 1, 2$ which are given by

$$\hat{\theta}_1^{(\neg \iota, 1)} | \hat{\nu}_1^{(2)} \sim N \left(\theta_1 + \rho \sqrt{\frac{(1 - \tau_\nu) \lambda \mathcal{I}_\nu}{(1 - \zeta)(\tau - \tilde{\tau}) \lambda \tilde{\mathcal{I}}}} (\hat{\nu}_1^{(2)} - \nu_1), \frac{1 - \rho^2}{(1 - \zeta)(\tau - \tilde{\tau}) \lambda \tilde{\mathcal{I}}} \right).$$

and

$$\hat{\theta}_2^{(\nu,1)} | \hat{\nu}_2^{(2)} \sim N \left(\theta_2 + \rho \sqrt{\frac{(1-\tau_\nu)(1-\lambda)\mathcal{I}_\nu}{(1-\zeta)(\tau-\tilde{\tau})(1-\lambda)\tilde{\mathcal{I}}}} (\hat{\nu}_2^{(2)} - \nu_2), \frac{1-\rho^2}{(1-\zeta)(\tau-\tilde{\tau})(1-\lambda)\tilde{\mathcal{I}}} \right).$$

Using these distributions we may simulate the remainder of observations from the pre-interim recruitment cohort at the interim analysis.

The post-interim recruitment cohort behaves in the same way as usual, if the trial continues in both sub-populations then the distributions are as those given in Section 3.6

$$\hat{\theta}_1^{(2)} \sim N(\theta_1, (\lambda(1-\tau)\tilde{\mathcal{I}})^{-1}), \hat{\theta}_2^{(2)} \sim N(\theta_2, ((1-\lambda)(1-\tau)\tilde{\mathcal{I}})^{-1}),$$

if the trial continues in only a single sub-population then the distribution for $i = 1, 2$

$$\hat{\theta}_i^{(2)} \sim N(\theta_i, ((1-\tau)\tilde{\mathcal{I}})^{-1}).$$

Hence we know all the necessary elements to optimise the interim decision of an adaptive enrichment trial design. While the details of the distributions for the estimates of $\boldsymbol{\theta}$ will not change in the overall hypothesis testing structure, the interim optimisation now uses an estimate for each θ with a reduced variance. We saw in Section 5.1 how the increased variance on the estimate of the treatment effect at the interim analysis reduced the overall performance of the Adaptive Enrichment trial, by introducing the secondary endpoint as we have we are able to compensate for this loss of information caused by the delayed response.

Example

Returning to the same setting we used in Section 5.1 we recall the common parameters,

$$\tilde{\mathcal{I}} = \left(\frac{\Phi^{-1}(0.9) + \Phi^{-1}(0.975)}{10} \right)^2,$$

$\lambda = 0.5$, $\tau = 0.5$ and we will use the extreme case where $\tilde{\tau} = 0.1$. In Section 5.1 we saw that when $\tilde{\tau} = 0.1$ there was a large reduction in the overall performance of the Adaptive Enrichment design. As usual we use the hypothesis testing methods described in Section 3.5.2.

Under the first formulation of the problem we use the gain function given in equation 4.5

$$G(\boldsymbol{\theta}, \kappa) = \lambda\theta_1\mathbb{I}(\mathcal{R}_1) + \theta_3\mathbb{I}(\mathcal{R}_3) + \theta_3\mathbb{I}(\mathcal{R}_b),$$

and under the second formulation of the problem we use the gain function from equation 4.6

$$G(\boldsymbol{\theta}, \kappa) = \lambda\theta_1\mathbb{I}(\mathcal{R}_1) + (1 - \lambda)\theta_2\mathbb{I}(\mathcal{R}_2) + (\lambda\theta_1 + (1 - \lambda)\theta_2)\mathbb{I}(\mathcal{R}_b).$$

To use comparable prior distributions, under the first formulation of the problem the marginal distributions for θ_1 and θ_2 should be

$$\theta_1 \sim N(9, 16), \theta_2 \sim N(3, 4)$$

and under the second formulation of the problem the marginal distributions for θ_1 and θ_2 should be

$$\theta_1 \sim N(6, 9), \theta_2 \sim N(6, 9).$$

Where θ_1 and θ_2 are independent.

We must also define how the short term endpoint will behave, we set $\tau_\nu = 0.25$, $\rho = 0.8$ and $\mathcal{I}_\nu = 0.75\tilde{\mathcal{I}}$. This choice of \mathcal{I}_ν suggests we expect less variability in the observations of the short term endpoint. Here we have chosen τ_ν and ρ to be the same for both sub-populations in either formulation of the problem, this is a very simplified assumption and something that may vary in practice. If we were to define different values of τ_ν and ρ for each sub-population we could still apply the same optimisation framework.

Formulation 1

Using the same correlation in the prior distribution as we do for the data we define the prior distributions to be

$$\begin{pmatrix} \theta_1 \\ \nu_1 \end{pmatrix} \sim N_2 \left(\begin{pmatrix} 9 \\ 6 \end{pmatrix} \begin{pmatrix} 16 & 10.1 \\ 10.1 & 10 \end{pmatrix} \right)$$

Interim analysis	% enriched	$\mathbb{E}_{\pi(\boldsymbol{\theta})}(G(\boldsymbol{\theta}, \kappa))$
Long term endpoint only	20.9%	3.66
Enhanced optimisation	41.9%	3.75
Immediate response	45.0%	3.88

Table 5.2: Comparing $\mathbb{E}_{\pi(\boldsymbol{\theta})}(G(\boldsymbol{\theta}, \kappa))$ of Bayes optimal Adaptive Enrichment design with a delayed response for formulation 1

and

$$\begin{pmatrix} \theta_2 \\ \nu_2 \end{pmatrix} \sim N_2 \left(\begin{pmatrix} 3 \\ 6 \end{pmatrix} \begin{pmatrix} 4 & 5.1 \\ 5.1 & 10 \end{pmatrix} \right)$$

which gives the same marginal distribution for the long term endpoints as we used previously. This is to say that we believe the treatment effect in the short term endpoint is consistent between the two sub-populations and we have the same amount of certainty about this in both sub-populations.

Table 5.2 shows how the Bayes optimal Adaptive Enrichment design performs when using the enhanced optimisation from the short term endpoint under the first formulation of the problem. We see that the enhanced optimisation has approximately doubled the probability of enriching the trial when compared to making the decision using only the long term endpoint, this brings this probability closer to that seen when assuming an immediate response. The expected gain of the enhanced Adaptive Enrichment design shows an increase over using only the long term endpoint, as with the probability of enriching the trial the expected gain has moved closer to the expected gain achieved by the trial using immediate response.

Formulation 2

Under the second formulation of the problem we use the prior distributions

$$\begin{pmatrix} \theta_1 \\ \nu_1 \end{pmatrix} \sim N_2 \left(\begin{pmatrix} 6 \\ 9 \end{pmatrix} \begin{pmatrix} 9 & 4.8 \\ 4.8 & 4 \end{pmatrix} \right)$$

Interim analysis	% enriched	$\mathbb{E}_{\pi(\theta)}(G(\theta, \kappa))$
Long term endpoint only	22.3%	2.07
Enhanced optimisation	56.6%	2.18
Immediate response	70.8%	2.44

Table 5.3: Comparing $\mathbb{E}_{\pi(\theta)}(G(\theta, \kappa))$ of Bayes optimal Adaptive Enrichment design with a delayed response for formulation 2

and

$$\begin{pmatrix} \theta_2 \\ \nu_2 \end{pmatrix} \sim N_2 \left(\begin{pmatrix} 6 \\ 9 \end{pmatrix} \begin{pmatrix} 9 & 4.8 \\ 4.8 & 4 \end{pmatrix} \right)$$

which as in the first formulation of the problem yields the marginal distributions for the long term endpoints as required. This particular prior suggests we expect the short term endpoint to show a stronger treatment effect than the long term endpoint, we have also assumed a lower prior variance suggesting we believe we know more about the short term endpoint prior to starting the trial.

In Table 5.3 we compare the enhanced optimisation with the optimisation based only on the long term endpoint to see how these methods compare with the immediate response version of the problem. We see a similar increase in the probability of enriching the trial as we did with the first formulation of the problem when using the enhanced optimisation compared to just using the long term endpoint, however this probability has not moved as close to the design assuming immediate response as we saw in the first formulation of the problem. There is a similar story for the expected gain of the designs, the enhanced Adaptive Enrichment design for delayed response has not recovered all of the expected gain that is achieved when using an immediate response.

Under both formulations of the problem we have seen that enhancing the optimisation procedure using the short term endpoint can provide a benefit. We see that the probability of enriching the trial for the enhanced optimisation procedure is closer to that given by a design assuming immediate response, from this change we see an improvement in the expected gain.

5.2.3 Adding to the joint model

In addition to the correlation structure we used in Section 5.2.1 the treatment effects for the short term endpoint and the long term endpoint may be directly related. We might assume that the treatment effect in the short term endpoint is simply some multiple of the treatment effect in the long term endpoint, for example

$$\nu_i = \xi\theta_i \text{ for } i = 1, 2.$$

If $\xi < 1$ then the treatment effect on the short term endpoint is smaller than what we ultimately expect to observe in the long term endpoint, if $\xi > 1$ the treatment effect is magnified for the short term endpoint. The prior distribution may now be updated using the observations of secondary endpoint directly at the interim analysis gives the posterior $\pi(\boldsymbol{\theta}|\boldsymbol{\nu})$, further reducing the uncertainty about the true treatment effect at the interim analysis.

5.2.4 Example

Returning to our example we experiment with an extreme case, we set $\tau_\nu = \tilde{\tau} = 0$ such that we have no observations of the primary endpoint at the interim analysis. We also set $\rho = 0$ meaning that the decision at the interim analysis is based solely on the posterior distribution given the estimate of the short term treatment effect. This choice of ρ is computationally simpler, we now simulate 1,000,000 values of $\hat{\nu}_1$ and $\hat{\nu}_2$, estimate the Bayes optimal decision and simulate a single realisation of the trial for each combination of $\hat{\nu}_1$ and $\hat{\nu}_2$. Finding the average gain of this simulation gives an estimate for the Bayes expected gain of the trial under this decision rule.

Using $\xi = 2$ and only the information from the short term endpoint at the interim analysis we apply the Bayes optimal decision rule under the first formulation of the problem. The probability of enriching for the Adaptive Enrichment trial is 26.1% and the expected gain is 3.69. Table 5.4 compares this with the alternatives, firstly we compare with the Adaptive Enrichment designs that always make the same decision from which we see that making an adaptation based only on the short term endpoint has provided a small improvement over the decision to never enrich that would have been made using only the prior

Decision rule	$\mathbb{E}_{\pi(\boldsymbol{\theta})}(G(\boldsymbol{\theta}, \kappa))$
Adaptive Enrichment always enrich	3.31
Adaptive Enrichment never enrich	3.62
Secondary endpoint only	3.69
Immediate response	3.88
Fixed sampling designs	$\mathbb{E}_{\pi(\boldsymbol{\theta})}(G(\boldsymbol{\theta}, \kappa))$
Multiple hypotheses	3.67
Fixed Enrichment	3.78

Table 5.4: Comparing $\mathbb{E}_{\pi(\boldsymbol{\theta})}(G(\boldsymbol{\theta}, \kappa))$ Adaptive Enrichment designs with no information on the final endpoint under formulation 1

distribution. Comparing the Adaptive Enrichment trials with the fixed sampling alternatives we see that the Adaptive Enrichment trial has provided a very small improvement over the fixed sampling design testing multiple hypotheses. The fixed Enrichment design gives a higher expected gain than this Adaptive Enrichment design, but if we were to increase ξ or \mathcal{I}_ν we would increase the overall performance of the trial.

Repeating this for the second formulation of the problem we see that the Bayes optimal Adaptive Enrichment trial using only the short term endpoint at the interim analysis is 43.2% giving an expected gain for the trial of 2.12. Comparing this with the alternative designs in table 5.5 we again see that an adaptive rule based only on a short term endpoint has provided a small improvement over the adaptive decision that would have been made using only the prior distribution. Comparing to the fixed sampling designs for this example the testing of multiple hypotheses was preferred and the Adaptive Enrichment design has matched the expected gain of this design.

The Adaptive Enrichment designs we have seen in this example do not provide any overall improvement over the fixed sampling alternatives, however that is not necessarily our aim if we are willing to make the decision based only on a short term endpoint. If we are unable to achieve an agreement on which fixed sampling design to use, despite the statistical argument for doing so, we have seen that starting recruitment for the trial and making the decision later based on some short term endpoint that is related to the final endpoint will not necessarily cause

Decision rule	$\mathbb{E}_{\pi(\boldsymbol{\theta})}(G(\boldsymbol{\theta}, \kappa))$
Adaptive enrichment always enrich	1.46
Adaptive enrichment never enrich	2.04
Secondary endpoint only	2.12
Immediate response	2.44
Fixed sampling designs	$\mathbb{E}_{\pi(\boldsymbol{\theta})}(G(\boldsymbol{\theta}, \kappa))$
Multiple hypotheses	2.12
Fixed enrichment	1.91

Table 5.5: Comparing $\mathbb{E}_{\pi(\boldsymbol{\theta})}(G(\boldsymbol{\theta}, \kappa))$ adaptive enrichment designs with no information on the final endpoint under formulation 2

a loss of performance for the trial.

5.3 Survival endpoints

5.3.1 A single survival endpoint

Up until this point all of our trial designs we have assumed that the observations from individuals are normally distributed, this has allowed us to simulate summary statistics that are also normally distributed. In the setting of clinical trials it is common to observe time to event or survival endpoints. We define a survival endpoint as the time after treatment until some event, this event could be death or some other form of disease progression, this is called the survival time. Collett (2015) gives a summary of appropriate techniques for the analysis of survival data.

Defining the log-rank score statistic

We take the approach of Turnbull and Jennison (2000), using the log-rank test for hypothesis testing in the setting of survival endpoints. Suppose for simplicity that we are using a single population for the trial, as we did in Chapter 2 we recruit n patients in total, n_1 of these patients are randomised to the new treatment and the remaining n_2 are randomised to the control. From these patients we observe the survival times which follow the hazard functions $h_A(t)$ for the new treatment and $h_B(t)$ for the control, from these we seek to test the null hypothesis $H_0 : h_1(t) = h_2(t)$ for all $t > 0$. The log-rank test is particularly

useful in the setting of proportional hazards,

$$h_1(t) = \xi h_2(t), \quad (5.9)$$

for some constant ξ .

The definition of the log-rank score statistic for K analyses, from Turnbull and Jennison (2000) Section 3.7 is as follows. Let d_k denote the number of uncensored events at analysis k . We denote the survival times of the patients by $t_{1,k} < t_{2,k} < \dots < t_{d_k,k}$ where each $t_{i,k}$ is measured as the time between study entry and failure. Let $r_{iA,k}$ and $r_{iB,k}$ be the numbers known to have survived for each treatment at the corresponding survival time $t_{i,k}$ and

$$\delta_{iB,k} = \begin{cases} 1 & \text{if the failure at time } k \text{ was on treatment } B \\ 0 & \text{if the failure at time } k \text{ was on treatment } A. \end{cases}$$

Then the log-rank score statistic at analysis k is

$$S_k = \sum_{i=1}^{d_k} \left(\delta_{iB,k} - \frac{r_{iB,k}}{r_{iA,k} + r_{iB,k}} \right) \quad (5.10)$$

For λ close to one and $\theta = \log(\xi)$ close to zero we use the approximation

$$S_k \sim N(\theta \mathcal{I}_k(\theta), \mathcal{I}_k(\theta)) \quad (5.11)$$

for sufficiently large information $\mathcal{I}_k(\theta)$ where $\mathcal{I}_k(\theta) \approx d_k/4$. Conditional on the observed information sequence the joint distribution of S_1, \dots, S_k approximates the standard form of joint distribution for score statistics given by Turnbull and Jennison (2000). Using the sequence S_1, \dots, S_k allows use to use the independent increments property, $S_1, S_2 - S_1$ independent, Whitehead (1997).

Adaptive Enrichment with survival endpoints

Returning to the Adaptive Enrichment design we may now consider how this might be implemented using a survival endpoint. We shall define the distribution of the log-rank score statistics in terms of $\tilde{\mathcal{I}}$ as we do for normally distributed

responses in Section 3.6. However in this setting splitting $\tilde{\mathcal{I}}$ is not as simple as simply splitting the number of patients between the sub-populations, we must make a number of assumptions about the data collection: we assume a proportional hazards model within each sub-population as in Equation 5.9; we assume that in each sub-population at each stage of the trial we may find the associated log-rank score statistic as in Equation 5.10 and that these log-rank score statistics each follow a distribution of the form given in Equation 5.11; and we assume that the recruitment and number of patients at risk and events in each sub-population are such that we observe the distributions of the log-rank score statistics that we shall define shortly. In practice this final assumption is highly unlikely as the number and times of each event that contribute to $\mathcal{I}_k(\theta_i)$ are random variables. At the end of the trial the exact distributions for the log-rank score statistics will be known and so this does not impact on the strong control of the FWER; the place where this does have an impact is in the optimisation of the interim decision, as this must be based on expected numbers of events.

Under these assumptions we may define the distributions of the log-rank score statistics for an Adaptive Enrichment design. Firstly the pre-interim recruitment cohort give log-rank score statistics over 2 analyses. At the interim analysis we observe log-rank score statistics from the events in the pre-interim recruitment cohort $\hat{S}_1^{(\iota)}$ in the first sub-population and $\hat{S}_2^{(\iota)}$ in the second sub-population where

$$\hat{S}_1^{(\iota)} \sim N(\lambda\tilde{\tau}\tilde{\mathcal{I}}\theta_1, \lambda\tilde{\tau}\tilde{\mathcal{I}}), \hat{S}_2^{(\iota)} \sim N((1-\lambda)\tilde{\tau}\tilde{\mathcal{I}}\theta_2, (1-\lambda)\tilde{\tau}\tilde{\mathcal{I}}).$$

For the rest of the trial we continue to follow up the pre-interim recruitment cohort, this gives the log-rank score statistics $\hat{S}_1^{(\neg\iota)}$ in the first sub-population and $\hat{S}_2^{(\neg\iota)}$ in the second sub-population, where

$$\hat{S}_1^{(\neg\iota)} \sim N(\lambda\tilde{\tau}\tilde{\mathcal{I}}\theta_1, \lambda\tilde{\tau}\tilde{\mathcal{I}}), \hat{S}_2^{(\neg\iota)} \sim N((1-\lambda)\tilde{\tau}\tilde{\mathcal{I}}\theta_2, (1-\lambda)\tilde{\tau}\tilde{\mathcal{I}}).$$

By independent increments of the score statistic at the end of the trial the log-rank score statistics for the pre-interim recruitment cohort are

$$\hat{S}_1^{(1)} = \hat{S}_1^{(\iota)} + \hat{S}_1^{(\neg\iota)}, \hat{S}_2^{(1)} = \hat{S}_2^{(\iota)} + \hat{S}_2^{(\neg\iota)}.$$

Under the first formulation of the problem we also find the log-rank score statistic in the full population $\hat{S}_3^{(1)} = \hat{S}_1^{(1)} + \hat{S}_2^{(1)}$. These log-rank score statistics for the pre-interim recruitment cohort follow the distributions,

$$\hat{S}_1^{(1)} \sim N(\lambda\tau\tilde{\mathcal{I}}\theta_1, \lambda\tau\tilde{\mathcal{I}}) \quad (5.12)$$

$$\hat{S}_2^{(1)} \sim N((1-\lambda)\tau\tilde{\mathcal{I}}\theta_2, (1-\lambda)\tau\tilde{\mathcal{I}}) \quad (5.13)$$

$$\begin{pmatrix} \hat{S}_1^{(1)} \\ \hat{S}_3^{(1)} \end{pmatrix} \sim N_2 \left(\begin{pmatrix} \lambda\tau\tilde{\mathcal{I}}\theta_1 \\ \tau\tilde{\mathcal{I}}\theta_3 \end{pmatrix}, \begin{pmatrix} \lambda\tau\tilde{\mathcal{I}} & \tau\tilde{\mathcal{I}} \\ \tau\tilde{\mathcal{I}} & \tau\tilde{\mathcal{I}} \end{pmatrix} \right).$$

The distributions of the log-rank score statistics for the post-interim recruitment cohort depend on which decision has been made, if we continue in both sub-populations in the sub-population 1 we have

$$\hat{S}_1^{(2)} \sim N(\lambda(1-\tau)\tilde{\mathcal{I}}\theta_1, \lambda(1-\tau)\tilde{\mathcal{I}}),$$

in sub-population 2 we have

$$\hat{S}_2^{(2)} \sim N((1-\lambda)(1-\tau)\tilde{\mathcal{I}}\theta_2, (1-\lambda)(1-\tau)\tilde{\mathcal{I}})$$

and under the first formulation of the problem the log-rank score statistic in the full population is $\hat{S}_3^{(2)} = \hat{S}_1^{(2)} + \hat{S}_2^{(2)}$ where

$$\begin{pmatrix} \hat{S}_1^{(2)} \\ \hat{S}_3^{(2)} \end{pmatrix} \sim N_2 \left(\begin{pmatrix} \lambda(1-\tau)\tilde{\mathcal{I}}\theta_1 \\ (1-\tau)\tilde{\mathcal{I}}\theta_3 \end{pmatrix}, \begin{pmatrix} \lambda(1-\tau)\tilde{\mathcal{I}} & (1-\tau)\tilde{\mathcal{I}} \\ (1-\tau)\tilde{\mathcal{I}} & (1-\tau)\tilde{\mathcal{I}} \end{pmatrix} \right).$$

The hypothesis testing may be conducted at the end of the trial in the same way as we described in Section 3.5.2, using the weighted inverse normal combination test to combine the P-value from the pre-interim and post-interim recruitment cohorts. By definition all follow up on patients recruited before the interim analysis contributes only to test statistics for the pre-interim recruitment cohort, as stated by Jenkins et al. (2011) this ensures independence between the test statistics of the pre-interim and post-interim recruitment cohorts as required for the combination tests.

Parallels to delayed response

We can clearly see a parallel with the delayed response from Section 5.1. The log-rank score statistics used to analyse the trial with a survival endpoint correspond exactly with estimates of the treatment effect since for $v_i^j = \text{var}(S_i^{(j)})$ for $i = 1, 2, 3$ and $j = 1, 2$

$$\hat{\theta}_i^{(j)} = \frac{S_i^{(j)}}{v_i^{(j)}}.$$

Using the same optimisation structure as Section 5.1 for Adaptive Enrichment trials using the log-rank score statistic for the survival endpoint will yield the same performance as achieved with delayed responses if all other parameters are the same, since the distributions are equivalent.

5.3.2 Using progression free survival to improve decision making at the interim analysis

In Section 5.2 we made use of a secondary endpoint to enhance the interim decision rule of the Adaptive Enrichment design in the setting of longitudinal data collection. We consider now how a similar approach could be taken in the context of survival data. For example suppose we collect times for both overall survival and progression free survival, where overall survival is the primary endpoint for the trial. Events for overall survival may occur slowly for overall survival compared to progression free survival, so at the interim analysis we may have more information on progression free survival.

If we construct the Adaptive Enrichment design using log-rank test based on overall survival we may construct the trial as described in Section 5.3.1; we remarked in Section 5.3.1 how this was equivalent to the delayed response seen in Section 5.1. In the same way we could extend the approach from Section 5.2 to this setting. Using the framework for enhancing the decision at the interim analysis of Adaptive Enrichment trials would maintain strong control of the FWER as the hypothesis testing structure would still only depend on overall survival.

To use the optimisation framework from Section 5.2 we would require a joint model for the log-rank score statistics of overall survival and progression free survival. In Section 5.2 we assumed that the distributions were bi-variate normal between the primary and secondary endpoint, if this is a reasonable assumption for overall survival and progression free survival we may proceed in exactly the same way. In Section 5.2.3 we extended the joint model further assuming the treatment effects were directly related, this is equivalent to assuming the log hazard ratios are directly related for overall and progression free survival; in this setting we might consider a relationship between the hazard ratios, perhaps we expect to see a stronger effect from the treatment on progression free survival than overall survival.

Chapter 6

Sensitivity of optimal designs to the model assumptions

6.1 Changing the prior distribution

Across all of our examples we have throughout Chapters 4 and 5 we have used the same prior distribution for each formulation of the problem. We have seen that Adaptive Enrichment designs can be an effective tool under these particular choices of the prior distribution. The framework for making the Bayes optimal decision introduced in Section 4.4 allows us to find the optimal decision under any choice of prior or gain function; with this in mind we investigate how Adaptive Enrichment designs perform under different choices for the prior distribution.

6.1.1 Returning to immediate response

In this Chapter we return to our assumption of an immediate response for all summary statistics while we investigate other aspects of the trial. The summary statistics follow the distributions given in Section 3.6 defined using $\tilde{\mathcal{I}}$, where $\tilde{\mathcal{I}}$ takes a fixed value across all trials. Under the first formulation of the problem we test the null hypotheses $H_{01} : \theta_1 \leq 0$ and $H_{03} : \theta_3 \leq 0$, and under the second formulation of the problem we test the null hypotheses $H_{01} : \theta_1 \leq 0$ and $H_{02} : \theta_2 \leq 0$. For the designs that use multiple populations λ is the proportion of the sample in the first sub-population and for Adaptive Enrichment designs τ is the proportion of the sample recruited in the pre-interim recruitment cohort.

In this Chapter we use the same fixed set of parameters defining the distributions of the observations we used throughout Chapters 4 and 5. These were: a fixed value $\tilde{\mathcal{I}}$,

$$\tilde{\mathcal{I}} = \left(\frac{\Phi^{-1}(0.9) + \Phi^{-1}(0.975)}{10} \right)^2$$

we set $\alpha = 0.025$ and use $\lambda = 0.5$ and $\tau = 0.5$ to govern the split of the sample between sub-populations and recruitment cohorts. As usual we will ensure strong control of the FWER as given in Equation 3.1 by using the hypothesis testing method described in Section 3.5.2

6.1.2 Alternative prior distributions

When changing the prior distributions one extreme case that still allows for comparison with our previous examples is to switch our choice of prior distribution for the first and second formulations of our problem.

Example: formulation 1

Under the first formulation of the problem we use the gain function given in equation 4.5,

$$G(\boldsymbol{\theta}, \kappa) = \lambda \theta_1 \mathbb{I}(\mathcal{R}_1) + \theta_3 \mathbb{I}(\mathcal{R}_3) + \theta_3 \mathbb{I}(\mathcal{R}_b)$$

and the new prior distribution

$$\theta_1 \sim N(6, 9), \theta_2 \sim N(6, 9) \tag{6.1}$$

where θ_1 and θ_2 are independent. Since $\lambda = 1/2$ we have $\theta_3 = \lambda \theta_1 + (1 - \lambda) \theta_2$ and may write the joint distribution as

$$\begin{pmatrix} \theta_1 \\ \theta_3 \end{pmatrix} \sim N_2 \left(\begin{pmatrix} 6 \\ 6 \end{pmatrix}, \begin{pmatrix} 9 & 4.5 \\ 4.5 & 4.5 \end{pmatrix} \right).$$

Table 6.1 shows the expected gain of the Bayes optimal Adaptive Enrichment trial for this example, comparing this with the expected gain from the comparable fixed sampling designs. We see that the fixed sampling design testing the null

Trial design	$\mathbb{E}_{\pi(\boldsymbol{\theta})}(G(\boldsymbol{\theta}, \kappa))$
Bayes optimal Adaptive Enrichment	3.33
Fixed Enrichment trial	1.91
Fixed sample multiple hypotheses	3.17

Table 6.1: Values of $\mathbb{E}_{\pi(\boldsymbol{\theta})}(G(\boldsymbol{\theta}, \kappa))$ with prior distribution from Equation 6.1 for formulation 1

hypotheses H_{01} and H_{03} is clearly the best fixed sampling design, the Adaptive Enrichment trial still offers an overall improvement in the Bayes expected gain of the trial. In fact this improvement is larger than the improvement we have previously observed from Adaptive Enrichment in the first formulation of the problem.

Example: formulation 2

Under the second formulation of the problem we use the gain function given in equation 4.6,

$$G(\boldsymbol{\theta}, \kappa) = \lambda\theta_1\mathbb{I}(\mathcal{R}_1) + (1 - \lambda)\theta_2\mathbb{I}(\mathcal{R}_2) + (\lambda\theta_1 + (1 - \lambda)\theta_2)\mathbb{I}(\mathcal{R}_b)$$

and the new prior distribution

$$\theta_1 \sim N(9, 16), \theta_2 \sim N(3, 4) \tag{6.2}$$

where θ_1 and θ_2 are independent.

Table 6.2 shows the expected gain of the Bayes optimal Adaptive Enrichment trial for this example, comparing this with the expected gain from the comparable fixed sampling designs. We see that the fixed Enrichment trial in the first sub-population is clearly the best choice. Although it is not the best overall the Adaptive Enrichment design provides an improvement over the fixed Enrichment design in the second sub-population and the fixed sample testing multiple hypotheses.

Trial design	$\mathbb{E}_{\pi(\boldsymbol{\theta})}(G(\boldsymbol{\theta}, \kappa))$
Bayes optimal Adaptive Enrichment	3.33
Fixed Enrichment trial sub-pop 1	3.77
Fixed Enrichment trial sub-pop 2	0.46
Fixed sample multiple hypotheses	2.82

Table 6.2: Values of $\mathbb{E}_{\pi(\boldsymbol{\theta})}(G(\boldsymbol{\theta}, \kappa))$ with prior distribution from Equation 6.2 for formulation 2

6.1.3 Uniform prior distributions

Another alternative would be to change the shape of the prior distribution altogether, a simple example of this would be to use a Uniform distribution instead of a Normal distribution.

Example: formulation 1

Recall that under this formulation of the problem we have typically used prior distributions given by

$$\theta_1 \sim N(9, 16), \theta_2 \sim N(3, 4)$$

where θ_1 and θ_2 are independent. Defining uniform priors in a similar area of the parameter space we use

$$\theta_1 \sim \text{Unif}(4, 12), \theta_2 \sim \text{Unif}(0, 8) \quad (6.3)$$

where again θ_1 and θ_2 are independent.

Under the uniform distribution the probability of enriching in the Bayes optimal Adaptive Enrichment trial is 35.7%, which is slightly different to our usual choices for normally distributed priors. Table 6.3 shows the expected gain of the Bayes optimal Adaptive Enrichment trial for this example, comparing this with the expected gain from the fixed sampling designs. We see that of the fixed sampling designs sampling in both sub-populations and testing both null hypotheses is now clearly the better choice. The Adaptive Enrichment trial provides a small improvement overall; the benefit here is smaller than we have previously seen for this formulation of the problem because the extreme values of $\boldsymbol{\theta}$, where Adaptive Enrichment is not as good as the fixed sampling alternatives,

Trial design	$\mathbb{E}_{\pi(\boldsymbol{\theta})}(G(\boldsymbol{\theta}, \kappa))$
Bayes optimal Adaptive Enrichment	3.30
Fixed Enrichment trial	3.01
Fixed sample multiple hypotheses	3.27

Table 6.3: Values of $\mathbb{E}_{\pi(\boldsymbol{\theta})}(G(\boldsymbol{\theta}, \kappa))$ for uniform priors from Equation 6.3 under formulation 1

are more likely under the uniform prior.

Example: formulation 2

Recall that under the second formulation of the problem we have typically used prior distributions given by

$$\theta_1 \sim N(6, 9), \theta_2 \sim N(6, 9).$$

where θ_1 and θ_2 are independent. Defining uniform priors in a similar area of the parameter space we use

$$\theta_1 \sim \text{Unif}(2, 10), \theta_2 \sim \text{Unif}(2, 10).$$

where again θ_1 and θ_2 are independent.

Under the uniform distribution the probability of enriching in the Bayes optimal Adaptive Enrichment trial is 64.5% which again is slightly different to what we have seen previously. Table 6.3 shows the expected gain of the Bayes optimal Adaptive Enrichment trial for this example, comparing this with the expected gain from the fixed sampling designs. We see that the fixed sampling trials are very similar in terms of their overall performance. The Adaptive Enrichment design again provides an improvement over the fixed sampling alternatives in this second formulation of the problem.

6.1.4 Comparing an alternative prior distribution

When considering the overall performance of the trial the same prior distributions for optimising the decision at the interim analysis and assessing the overall performance of the trials. In practice the prior distribution used for optimising

Trial design	$\mathbb{E}_{\pi(\boldsymbol{\theta})}(G(\boldsymbol{\theta}, \kappa))$
Bayes optimal Adaptive Enrichment	2.09
Fixed Enrichment trial	1.78
Fixed sample multiple hypotheses	1.83

Table 6.4: Comparing $\mathbb{E}_{\pi(\boldsymbol{\theta})}(G(\boldsymbol{\theta}, \kappa))$ for competing trial designs for formulation 2

the Adaptive Enrichment will be a summary of the prior beliefs of the investigators; however there may not be full agreement on this prior.

Suppose we have one prior distribution that summarises the entire beliefs of the trial team, $\pi_a(\boldsymbol{\theta})$ say, we use this prior to optimise the Adaptive Enrichment trial and find the overall performance in the usual way. Perhaps we have one member of the development team who wishes to assess the design under their own prior beliefs, $\pi_b(\boldsymbol{\theta})$ say.

Example: formulation 1

Under the first formulation of the problem we define the prior distribution summarising the overall beliefs $\pi_a(\boldsymbol{\theta})$ to be

$$\theta_1 \sim N(9, 16), \theta_2 \sim N(3, 4).$$

where θ_1 and θ_2 are independent. Since $\lambda = 1/2$ and $\theta_3 = \lambda\theta_1 + (1 - \lambda)\theta_2$ this is equivalent to

$$\begin{pmatrix} \hat{\theta}_1 \\ \hat{\theta}_3 \end{pmatrix} \sim N_2 \left(\begin{pmatrix} 9 \\ 6 \end{pmatrix}, \begin{pmatrix} 16 & 8 \\ 8 & 5 \end{pmatrix} \right).$$

We define the alternative prior distribution $\pi_b(\boldsymbol{\theta})$ to be

$$\theta_1 \sim N(6, 9), \theta_2 \sim N(6, 9),$$

where θ_1 and θ_2 are independent. Since $\lambda = 1/2$ we have $\theta_3 = \lambda\theta_1 + (1 - \lambda)\theta_2$ we may write the joint distribution as

$$\begin{pmatrix} \hat{\theta}_1 \\ \hat{\theta}_3 \end{pmatrix} \sim N_2 \left(\begin{pmatrix} 6 \\ 6 \end{pmatrix}, \begin{pmatrix} 9 & 4.5 \\ 4.5 & 4.5 \end{pmatrix} \right).$$

Trial design	$\mathbb{E}_{\pi_b(\boldsymbol{\theta})}(G(\boldsymbol{\theta}, \kappa))$
Adaptive Enrichment optimal for $\pi_a(\boldsymbol{\theta})$	2.82
Fixed Sample testing multiple hypotheses	3.17
Fixed Enrichment	1.91
Adaptive Enrichment optimal for $\pi_b(\boldsymbol{\theta})$	3.33

Table 6.5: Comparing $\mathbb{E}_{\pi(\boldsymbol{\theta})}(G(\boldsymbol{\theta}, \kappa))$ for different prior distributions under formulation 1

Trial design	$\mathbb{E}_{\pi_b(\boldsymbol{\theta})}(G(\boldsymbol{\theta}, \kappa))$
Adaptive Enrichment optimal for $\pi_a(\boldsymbol{\theta})$	2.05
Fixed Sample testing multiple hypotheses	2.12
Fixed Enrichment in sub-population 1	1.91
Fixed Enrichment in sub-population 2	1.91
Adaptive Enrichment optimal for $\pi_b(\boldsymbol{\theta})$	2.44

Table 6.6: Comparing $\mathbb{E}_{\pi(\boldsymbol{\theta})}(G(\boldsymbol{\theta}, \kappa))$ for different prior distributions under formulation 2

In Table 6.5 we see a comparison of the expected gain under $\pi_b(\boldsymbol{\theta})$. We see that when the interim decision of the Adaptive Enrichment design is optimised for $\pi_a(\boldsymbol{\theta})$ then $\mathbb{E}_{\pi_b(\boldsymbol{\theta})}(G(\boldsymbol{\theta}, \kappa))$ is lower than the best fixed sampling alternative; if we were to optimise for $\pi_b(\boldsymbol{\theta})$ then the Adaptive Enrichment design is the best choice of trial.

Example: formulation 2

We make the equivalent analysis in formulation 2 using the same choices for $\pi_a(\boldsymbol{\theta})$ and $\pi_b(\boldsymbol{\theta})$. We see in Table 6.6 that we see the same outcomes as we did for formulation 1. When the interim decision of the Adaptive Enrichment design is optimised for $\pi_a(\boldsymbol{\theta})$ then $\mathbb{E}_{\pi_b(\boldsymbol{\theta})}(G(\boldsymbol{\theta}, \kappa))$ is lower than the best fixed sampling alternative; if we were to optimise for $\pi_b(\boldsymbol{\theta})$ then the Adaptive Enrichment design is the best choice of trial.

The examples in this section demonstrate the importance of the choice of prior distribution when evaluating the performance of Adaptive Enrichment designs. We have shown that when optimising the interim decision of Adaptive Enrichment designs the choice of prior will impact on the overall performance of the design;

this also shows that using a simple rule without knowledge of the Bayes optimal decision boundary may be a very bad choice. Under the first formulation of the problem in our examples the Adaptive Enrichment design is always the best choice of design, however the results for the second formulation of the problem demonstrate that under some prior distributions the fixed sampling alternatives offer a better overall performance than Adaptive Enrichment designs; the consequence of this is that in practice we know that the Adaptive Enrichment design will not always be the best choice, the overall performance should be considered before choosing to use an Adaptive Enrichment design over the fixed sampling alternatives.

6.2 Changing the parameters of the design

In addition to assuming the same prior distribution throughout Chapters 4 and 5 we have also assumed a fixed pattern of recruitment. We took the proportion of the sample in the first sub-population to be half of the overall sample, setting $\lambda = 1/2$. Similarly we set $\tau = 1/2$ conducting the interim analysis after half of the total sample had been collected. The final parameter governing recruitment that we have used a single value for is the fixed value for $\tilde{\mathcal{I}}$ which we set to be

$$\tilde{\mathcal{I}} = \left(\frac{\Phi^{-1}(0.9) + \Phi^{-1}(0.975)}{10} \right)^2.$$

We now experiment with different choices for the values of these parameters.

6.2.1 Sub-population proportion

The choice of the proportion of the population in the first sub-population (given by λ) will vary depending on the criteria for being a member of the sub-population. Typically the sub-population may be selected by a predictive biomarker indicating that a patient may respond in a particular way to the new treatment, and so we may not be able to select the proportion of patients that will be recruited in this sub-population. The methods presented in Chapters 4 and 5 do not rely on a specific choice of λ ; we may conduct the trial and optimise the decision at the interim analysis for all $\lambda \in (0, 1)$.

Example

Taking all other parameters as we have in our previous examples under the first formulation of the problem we use the familiar prior distribution

$$\theta_1 \sim N(9, 16), \theta_2 \sim N(3, 4),$$

where θ_1 and θ_2 are independent; the joint distribution of θ_1 and θ_3 will vary with λ since $\theta_3 = \lambda\theta_1 + (1 - \lambda)\theta_2$. The gain function for the first formulation of the problem given in equation 4.5 is

$$G(\boldsymbol{\theta}, \kappa) = \lambda\theta_1\mathbb{I}(\mathcal{R}_1) + \theta_3\mathbb{I}(\mathcal{R}_3) + \theta_3\mathbb{I}(\mathcal{R}_b).$$

For the second formulation of the problem we define the prior distributions to be

$$\theta_1 \sim N(6, 9), \theta_2 \sim N(6, 9).$$

where θ_1 and θ_2 are independent and use the gain function given in equation 4.6,

$$G(\boldsymbol{\theta}, \kappa) = \lambda\theta_1\mathbb{I}(\mathcal{R}_1) + (1 - \lambda)\theta_2\mathbb{I}(\mathcal{R}_2) + (\lambda\theta_1 + (1 - \lambda)\theta_2)\mathbb{I}(\mathcal{R}_b).$$

Varying the value of λ we apply the optimised decision rule and find the Bayes expected gain of the Adaptive Enrichment trial and compare this with the Bayes expected gain of each of the fixed sampling alternatives in Table 6.7. We see that under the first formulation of the problem the Adaptive Enrichment trial gives a small performance increase over both fixed sampling designs over many choices of λ . We note that as λ increases we also see that the fixed Enrichment trial improves when compared with the fixed sample testing multiple hypotheses.

Under the second formulation of the problem we evaluate values of λ up to 0.5 since we have symmetry between the sub-populations. We see that for low values of λ (and hence a high proportion of the sample in the second sub-population) the fixed Enrichment trial in the second sub-population is the best choice of trial; both fixed Enrichment designs will give the same probability of rejecting the null hypothesis by symmetry, the difference is because of the appearance of λ in the gain function. For $\lambda = 0.4$ and 0.5 the Adaptive Enrichment trial gives

	Formulation 1				
	$\lambda = 0.1$	$\lambda = 0.3$	$\lambda = 0.5$	$\lambda = 0.7$	$\lambda = 0.9$
Adaptive Enrichment	1.06	2.39	3.88	5.25	6.85
Fixed Enrichment	0.75	2.27	3.77	5.28	6.79
Multiple hypotheses	1.03	2.24	3.74	5.25	6.70

	Formulation 2				
	$\lambda = 0.1$	$\lambda = 0.2$	$\lambda = 0.3$	$\lambda = 0.4$	$\lambda = 0.5$
Adaptive Enrichment	2.93	2.73	2.56	2.46	2.44
Fixed Enrichment sub-pop 1	0.38	0.76	1.14	1.52	1.91
Fixed Enrichment sub-pop 2	3.42	3.04	2.66	2.28	1.91
Multiple hypotheses	2.89	2.54	2.31	2.17	2.12

Table 6.7: Comparing $\mathbb{E}_{\pi(\theta)}(G(\theta, \kappa))$ varying values of λ

the highest Bayes expected gain. Comparing the Bayes expected gain of the fixed sampling design testing multiple hypotheses to the Bayes optimal Adaptive Enrichment design we see that the advantage of the Adaptive Enrichment trial increases as λ gets closer to 0.5.

6.2.2 Timing of the interim analysis

The timing of the interim analysis dictates both how much of the sample has been recruited and hence how much information we have when making the decision at the interim analysis, and how much of the sample we can alter the recruitment for. We have used $\tau = 1/2$ previously, but this is not the only option. In practice we might alter the timing of the interim analysis to improve the overall performance of the trial. Whenever the interim analysis is to be conducted, we are able to optimise the decision using the Bayes optimal decision introduced in Section 4.4 since the distributions of the remaining observations are known.

Example

Returning to our example we now set $\lambda = 1/2$ and allow τ to vary. Table 6.8 shows how the Bayes optimal Adaptive Enrichment trial performs for each formulation of the problem. We see that under the both formulations of the problem there would be an advantage to conducting the interim analysis earlier in the trial. This may seem strange since the decision will be made with more noise, however the advantage is that more of the sample may be influenced by the

Formulation 1	
τ	$\mathbb{E}_{\pi(\theta)}(G(\theta, \kappa))$
0.25	3.94
0.50	3.88
0.75	3.80

Formulation 2	
τ	$\mathbb{E}_{\pi(\theta)}(G(\theta, \kappa))$
0.25	2.52
0.50	2.44
0.75	2.22

Table 6.8: Comparing $\mathbb{E}_{\pi(\theta)}(G(\theta, \kappa))$ varying values of τ

interim decision. Using this knowledge we could optimise the timing of the interim analysis for any particular Adaptive Enrichment trial, evaluating the expected gain of the Bayes optimal Adaptive Enrichment trials for different values of τ to find the value of τ that maximises the Bayes expected gain.

6.2.3 Total sample size of the trial

We fixed $\tilde{\mathcal{I}}$ for each trial to ensure that the Adaptive Enrichment design remained comparable in this aspect to the fixed sampling designs, our choice of

$$\tilde{\mathcal{I}} = \left(\frac{\Phi^{-1}(0.9) + \Phi^{-1}(0.975)}{10} \right)^2$$

was to give power of $1 - \beta = 0.9$ testing $H_0 : \theta \leq 0$ at $\alpha = 0.025$ when $\hat{\theta} \sim N(\theta, \tilde{\mathcal{I}})$. However this nominal power level is based on the sample size we are able to achieve and this may well be chosen to also balance the cost of the trial. Given that the sample size will not always achieve this value of $\tilde{\mathcal{I}}$ we might wish to know how Adaptive Enrichment will perform as $\tilde{\mathcal{I}}$ varies.

Example

Returning once more to our example from this Chapter we set $\lambda = 1/2$ and $\tau = 1/2$,

$$\tilde{\mathcal{I}} = \left(\frac{\Phi^{-1}(1 - \beta) + \Phi^{-1}(0.975)}{10} \right)^2$$

where we vary the nominal power $1 - \beta$.

Formulation 1					
$1 - \beta$		0.70	0.80	0.90	0.95
Adaptive Enrichment		2.87	3.32	3.88	4.26
Fixed Enrichment		3.15	3.46	3.77	3.95
Multiple hypotheses		2.65	3.13	3.74	4.17

Formulation 2					
$1 - \beta$		0.70	0.80	0.90	0.95
Adaptive Enrichment		1.65	1.98	2.44	2.75
Fixed Enrichment		1.30	1.62	1.91	2.10
Multiple hypotheses		1.39	1.64	2.12	2.53

Table 6.9: Comparing $\mathbb{E}_{\pi(\boldsymbol{\theta})}(G(\boldsymbol{\theta}, \kappa))$ varying values of $1 - \beta$

Table 6.9 shows how the Bayes expected gain of the Bayes optimal Adaptive Enrichment trial compares with the alternative fixed sampling designs as the nominal power $1 - \beta$ varies for each formulation of the problem. Under the first formulation of the problem we see that as $1 - \beta$ increases the Bayes expected gain of the Bayes optimal Adaptive Enrichment increases in comparison to the fixed Enrichment trial, the adaptive design has a lower performance than the fixed Enrichment design for lower values of $1 - \beta$ and is higher for $1 - \beta = 0.9$. Conversely comparing the Bayes expected gain of the adaptive design with the fixed design testing multiple hypotheses we see that while the performance of the fixed design is always lower, the advantage of the Adaptive Enrichment design decreases as $1 - \beta$ increases.

Under the second formulation of the problem we see that the Bayes optimal Adaptive Enrichment trial maintains a similar overall advantage in the Bayes expected gain to the fixed sampling designs. For all values of $1 - \beta$ the fixed trial testing multiple hypotheses is preferred to the fixed Enrichment trial, as $1 - \beta$ increases the advantage of testing multiple hypotheses increases. If we think about the typical shape of a power curve this is due to the fact that the fixed enrichment trial is in a shallower area of the curve than the designs testing multiple hypotheses.

Chapter 7

Variations to the Adaptive Enrichment design

7.1 Hypothesis testing methodology

The methods we have used to conduct the hypothesis tests across all examples in Chapters 4, 5 and 6 are only one possible example of conducting hypothesis test suitable for the trials we have considered. We showed in Section 3.2.3 that all hypothesis testing procedures that ensure strong control of the FWER can be written as closed testing procedures, within this we have used Simes method to test the intersection hypothesis. However, this is not the only method we could have chosen. Similarly the weighted inverse normal was not the only possible choice of combination test we could have made. We could use an alternative method for the overall hypothesis testing procedure we use in the Adaptive Enrichment trial.

Once again in this Chapter we will use the same standard setup where using the distributions set out in Section 3.6 where we choose \tilde{I} for all trials to be

$$\tilde{I} = \left(\frac{\Phi^{-1}(0.9) + \Phi^{-1}(0.975)}{10} \right)^2,$$

setting $\alpha = 0.025$, $\lambda = 0.5$ and $\tau = 0.5$. For both formulations of the problem we must set the method for hypothesis testing.

Under the first formulation of the problem we use the gain function from equation 4.5

$$G(\boldsymbol{\theta}, \kappa) = \lambda\theta_1\mathbb{I}(\mathcal{R}_1) + \theta_3\mathbb{I}(\mathcal{R}_3) + \theta_3\mathbb{I}(\mathcal{R}_b),$$

and the prior distributions

$$\theta_1 \sim N(9, 16), \theta_2 \sim N(3, 4).$$

where θ_1 and θ_2 are independent. Since $\lambda = 1/2$ we have $\theta_3 = \lambda\theta_1 + (1 - \lambda)\theta_2$ and may write the joint distribution as

$$\begin{pmatrix} \hat{\theta}_1 \\ \hat{\theta}_3 \end{pmatrix} \sim N_2 \left(\begin{pmatrix} 9 \\ 6 \end{pmatrix}, \begin{pmatrix} 16 & 8 \\ 8 & 5 \end{pmatrix} \right).$$

Under the second formulation of the problem we use the gain function from equation 4.6

$$G(\boldsymbol{\theta}, \kappa) = \lambda\theta_1\mathbb{I}(\mathcal{R}_1) + (1 - \lambda)\theta_2\mathbb{I}(\mathcal{R}_2) + (\lambda\theta_1 + (1 - \lambda)\theta_2)\mathbb{I}(\mathcal{R}_b).$$

For the second formulation of the problem we chose the symmetric prior

$$\theta_1 \sim N(6, 9), \theta_2 \sim N(6, 9).$$

where θ_1 and θ_2 are independent

7.1.1 Improved Simes method for the intersection hypothesis

Under the first formulation of the problem where we test the null hypotheses as $H_{01} : \theta_1 \leq 0$ and $H_{03} : \theta_3 \leq 0$ we know that there is a correlation between the estimates of the treatment effect between the sub-population of interest and the full population, since patients in the sub-population are also in the full population. Given λ and under the assumption of independence between the sub-population and the complement we know this correlation is $\rho = \sqrt{\lambda}$. This correlation will cause Simes method to become conservative for testing the intersection hypothesis.

λ	ρ	$FWER$	% of nominal α
0.25	0.50	0.024	96.0
0.50	0.71	0.023	91.0
0.75	0.87	0.021	84.8
0.90	0.95	0.020	81.9

Table 7.1: Conservatism of Simes method for the intersection hypothesis given positive correlation between variables

This knowledge can be used to improve Simes method, recall from equation 3.4 that the intersection p-value is given by $P_{12} = \min(2\min(P_1, P_2), \max(P_1, P_2))$. Since $\max(P_1, P_2)$ will reject the intersection hypothesis when the individual tests also reject the individual null hypotheses at level α , in order to maintain consonance we will replace $2\min(P_1, P_2)$ in the formula by $\psi_\rho \min(P_1, P_2)$. We choose ψ_ρ to spend the full α for a given ρ . Thus the improved Simes method for the intersection p-value is given by,

$$P_{12} = \min(\psi_\rho \min(P_1, P_2), \max(P_1, P_2)). \quad (7.1)$$

Example

We can find exactly how conservative Simes method is in terms of the FWER, Table 7.1 shows how this conservatism increases with λ in the fixed sampling design testing multiple hypotheses. For the fixed sampling design comparing multiple hypotheses we compare the effectiveness of the improved Simes method from equation 7.1 where for our choice of $\lambda = 0.5$ we find $\psi_\rho = 1.77$ for $\rho = \sqrt{0.5}$, with the original version of Simes method from Equation 3.4 in Table 7.2. We see that the improved version of Simes method gives a small increase in the probability of rejecting the null hypotheses in all cases, increasing the corresponding expected gain.

Using the improved Simes method given by equation 7.1 we find the Bayes expected gain of the Bayes optimal Adaptive Enrichment design and compare this with the expected gain of the fixed sampling alternatives in table 7.3. This shows that the improved version of Simes method gives a higher Bayes expected gain for both the Adaptive Enrichment trial and the fixed sampling trial testing multiple hypotheses. The improvement in the Adaptive Enrichment design is

Original Simes					
θ_1	θ_3	$\mathbb{P}_{\boldsymbol{\theta}}(\mathcal{R}_1)$	$\mathbb{P}_{\boldsymbol{\theta}}(\mathcal{R}_3)$	$\mathbb{P}_{\boldsymbol{\theta}}(\mathcal{R}_b)$	$\mathbb{E}_{\boldsymbol{\theta}}(G(\boldsymbol{\theta}, \kappa))$
10	10	0.01	0.24	0.61	8.59
10	5	0.21	0.02	0.33	2.82
10	0	0.49	0.00	0.02	2.48
0	0	0.01	0.01	0.01	0.00

Improved Simes					
θ_1	θ_3	$\mathbb{P}_{\boldsymbol{\theta}}(\mathcal{R}_1)$	$\mathbb{P}_{\boldsymbol{\theta}}(\mathcal{R}_2)$	$\mathbb{P}_{\boldsymbol{\theta}}(\mathcal{R}_b)$	$\mathbb{E}_{\boldsymbol{\theta}}(G(\boldsymbol{\theta}, \kappa))$
10	10	0.01	0.25	0.62	8.67
10	5	0.23	0.02	0.33	2.90
10	0	0.51	0.00	0.02	2.57
0	0	0.01	0.01	0.01	0.00

Table 7.2: Comparing Simes and Improved Simes under formulation 1: \mathcal{R}_1 is the event that we only reject H_{01} , \mathcal{R}_3 is the event that we only reject H_{03} and \mathcal{R}_b is the event that we reject both H_{01} and H_{03}

higher than the improvement in the fixed sampling trial, and we now see that the advantage of the Adaptive Enrichment design is higher than we had using the original version of Simes method. This shows that the improved version of Simes method can provide an improvement for any clinical trial under the first formulation of the problem and is particularly beneficial for the Adaptive Enrichment design.

7.1.2 Dunnet type rule for the intersection hypothesis

A second method that would allow us to make use of the correlation between the sub-population and the full population uses the principle introduced by Dunnett (1955). This method makes use of the correlation introduced when using a common control arm for multiple treatments, while we do not have a common control we do know the joint distribution of the summary statistics. Under the first formulation of the problem we are testing the null hypotheses $H_{01} : \theta_1 \leq 0$ and $H_{03} : \theta_3 \leq 0$, when conducting a fixed sampling trial to test these hypotheses we find Z-values Z_1 and Z_3 . When $\theta_1 = \theta_3 = 0$

$$\begin{pmatrix} Z_1 \\ Z_3 \end{pmatrix} \sim \left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \right)$$

Original Simes	
Trial design	$\mathbb{E}_{\pi(\boldsymbol{\theta})}(G(\boldsymbol{\theta}, \kappa))$
Adaptive Enrichment	3.88
Fixed Enrichment	3.78
Fixed sample multiple hypotheses	3.74

Improved Simes	
Trial design	$\mathbb{E}_{\pi(\boldsymbol{\theta})}(G(\boldsymbol{\theta}, \kappa))$
Adaptive Enrichment	3.94
Fixed sample multiple hypotheses	3.80

Table 7.3: Comparing $\mathbb{E}_{\pi(\boldsymbol{\theta})}(G(\boldsymbol{\theta}, \kappa))$ using Improved Simes method for the intersection hypothesis under formulation 1

and we reject $H_{01} \cap H_{03}$ when $\max(Z_1, Z_3) > r$ where r is such that

$$\mathbb{P}(\max(Z_1, Z_3) > r) \leq \alpha$$

We may find the corresponding p-value under the Dunnett type method, for observations z_1 and z_3 we define

$$d = \max(z_1, z_3)$$

the p-value for the intersection hypothesis is given by

$$p_{13} = \mathbb{P}(Z_1 \geq d \cup Z_3 \geq d). \quad (7.2)$$

Defining the test in this way ensures that the testing procedure is consonant as defined in Section 3.2.4.

Example

As we did with the improved Simes method we start by comparing the Dunnett method from equation 7.2 with the original version of Simes method from equation 3.4 for the fixed sampling design comparing multiple hypotheses, Table 7.4 shows this comparison, we see that the Dunnett type method performs very similarly to the improved Simes method, showing a small improvement over Simes method when $\theta_1 = 10$ and $\theta_3 = 0$.

		Original Simes			
θ_1	θ_3	$\mathbb{P}_{\boldsymbol{\theta}}(\mathcal{R}_1)$	$\mathbb{P}_{\boldsymbol{\theta}}(\mathcal{R}_3)$	$\mathbb{P}_{\boldsymbol{\theta}}(\mathcal{R}_b)$	$\mathbb{E}_{\boldsymbol{\theta}}(G(\boldsymbol{\theta}, \kappa))$
10	10	0.01	0.24	0.61	8.59
10	5	0.21	0.02	0.33	2.82
10	0	0.49	0.00	0.02	2.48
0	0	0.01	0.01	0.01	0.00

		Improved Simes			
θ_1	θ_3	$\mathbb{P}_{\boldsymbol{\theta}}(\mathcal{R}_1)$	$\mathbb{P}_{\boldsymbol{\theta}}(\mathcal{R}_2)$	$\mathbb{P}_{\boldsymbol{\theta}}(\mathcal{R}_b)$	$\mathbb{E}_{\boldsymbol{\theta}}(G(\boldsymbol{\theta}, \kappa))$
10	10	0.01	0.25	0.62	8.67
10	5	0.23	0.02	0.33	2.90
10	0	0.51	0.00	0.02	2.57
0	0	0.01	0.01	0.01	0.00

		Dunnett			
θ_1	θ_3	$\mathbb{P}_{\boldsymbol{\theta}}(\mathcal{R}_1)$	$\mathbb{P}_{\boldsymbol{\theta}}(\mathcal{R}_2)$	$\mathbb{P}_{\boldsymbol{\theta}}(\mathcal{R}_b)$	$\mathbb{E}_{\boldsymbol{\theta}}(G(\boldsymbol{\theta}, \kappa))$
10	10	0.01	0.25	0.62	8.67
10	5	0.23	0.02	0.33	2.89
10	0	0.52	0.00	0.02	2.60
0	0	0.01	0.01	0.01	0.00

Table 7.4: Comparing intersection hypothesis tests under formulation 1: \mathcal{R}_1 is the event that we only reject H_{01} , \mathcal{R}_3 is the event that we only reject H_{03} and \mathcal{R}_b is the event that we reject both H_{01} and H_{03} .

Original Simes	
Trial design	$\mathbb{E}_{\pi(\boldsymbol{\theta})}(G(\boldsymbol{\theta}, \kappa))$
Adaptive Enrichment	3.88
Fixed Enrichment	3.79
Fixed sample multiple hypotheses	3.67
Improved Simes	
Trial design	$\mathbb{E}_{\pi(\boldsymbol{\theta})}(G(\boldsymbol{\theta}, \kappa))$
Adaptive Enrichment	3.94
Fixed sample multiple hypotheses	3.80
Dunnett	
Trial design	$\mathbb{E}_{\pi(\boldsymbol{\theta})}(G(\boldsymbol{\theta}, \kappa))$
Adaptive Enrichment	3.96
Fixed sample multiple hypotheses	3.77

Table 7.5: Comparing $\mathbb{E}_{\pi(\boldsymbol{\theta})}(G(\boldsymbol{\theta}, \kappa))$ using improved tests for the intersection hypothesis

Using the Dunnett type method we find the Bayes expected gain of the Bayes optimal Adaptive Enrichment design and compare this with the expected gain of the fixed sampling alternatives in Table 7.5 we see that using the Dunnett type method gives an overall improvement over the original version of Simes method for both the fixed sampling and Adaptive designs. When compared with the improved version of Simes method we see that for the fixed sampling design the improved version of Simes method gives a higher Bayes expected gain, however for the adaptive design we see that the Dunnett type method is marginally better. There is no clearly preferred method between the improved Simes and Dunnett type methods, but in both cases we see that we can improve on the basic hypothesis testing structure we used throughout Chapters 4, 5 and 6.

7.1.3 χ^2 combination test

The weighted inverse normal described in Section 3.4.1 is known to be the optimal choice of combination test when the weights are chosen in proportion to the sample size within each stage of the trial. A commonly used alternative is the χ^2 combination test introduced by Fisher (1925). Suppose we have p-values

$P^{(1)}$ and $P^{(2)}$ from the first and second stages of the trial then under the null hypothesis

$$-2\log(P^{(1)}P^{(2)}) \sim \chi_4^2.$$

Suppose we have a random variable R following the same distribution so $R \sim \chi_4^2$, the combined p-value for testing the null hypothesis at the end of the trial is given by

$$p^{(c)} = \mathbb{P}(R \geq -2\log(p^{(1)}, p^{(2)})) \quad (7.3)$$

where $p^{(1)}$ and $p^{(2)}$ are the p-values observed in each stage of the trial.

Example

Consider a trial testing only $H_{03} : \theta_3 \leq 0$ where we conduct an interim analysis that always continues in the full population, but suppose we do not know the proportion of the sample τ that has been used at the interim analysis. From the pre-interim recruitment cohort we observe

$$\hat{\theta}_3^{(1)} \sim N(10, (\tau\tilde{I})^{-1})$$

and from the post-interim recruitment cohort we observe

$$\hat{\theta}_3^{(2)} \sim N(10, ((1 - \tau)\tilde{I})^{-1}).$$

Choosing weights $w_1 = w_2 = 0.5$ we compare the weighted inverse normal combination test with inverse χ^2 combination test from equation 7.3 as τ varies in Figure 7-1. We see that as τ varies the probability of rejecting the null hypothesis with the χ^2 combination test varies less, the probability of rejecting the null hypothesis under the weighted inverse normal is higher for choices of τ close to the weights but decreases the further from optimal the weights become. We have typically used $\tau = 0.5$ meaning that the sample is either split with half of the sample in the first and second recruitment cohorts, or a quarter in the first cohort and three quarters of the sample in the second cohort. We see that in both of these cases that the weighted inverse normal combination test is more powerful, thus if the proportion of the sample in each stage of the trial is well understood the weighted inverse normal will be the better choice (it would even be possible

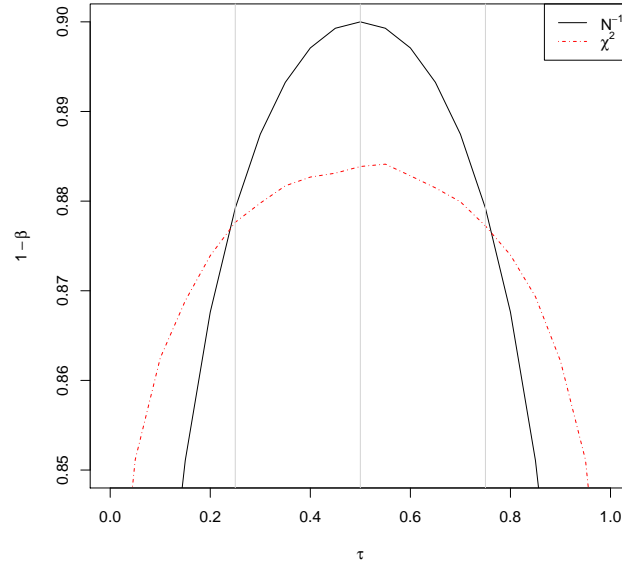


Figure 7-1: Comparing the power of combination tests

to optimise the weights to account for the adaptation).

Using the χ^2 combination test we may still find the Bayes optimal Adaptive Enrichment design, Table 7.6 compares the Bayes expected gain of the Bayes optimal Adaptive Enrichment designs using the weighted inverse normal and χ^2 combination tests for each formulation of our problem. Under both formulations of the problem we see that unsurprisingly the Bayes expected gain of the Bayes optimal Adaptive Enrichment trials is much lower when using the χ^2 combination test.

7.1.4 Changing the ordering of testing methods

The final element of the hypothesis testing that we kept consistent throughout all examples of Adaptive Enrichment trials in Chapters 2, 4 and 5 is the ordering of the overall testing procedure. We have found an intersection p-value within each stage of the trial and then used a combination test to find overall P-values for the individual and intersection hypothesis. This has the advantage that when the trial is enriched at the interim analysis we may set the intersection P-value

Formulation 1	
Combination test	$\mathbb{E}_{\pi(\boldsymbol{\theta})}(G(\boldsymbol{\theta}, \kappa))$
Weighted inverse normal	3.88
Inverse χ^2	3.25

Formulation 2	
Combination test	$\mathbb{E}_{\pi(\boldsymbol{\theta})}(G(\boldsymbol{\theta}, \kappa))$
Weighted inverse normal	2.44
Inverse χ^2	2.26

Table 7.6: Comparing $\mathbb{E}_{\pi(\boldsymbol{\theta})}(G(\boldsymbol{\theta}, \kappa))$ using different combination tests under each formulation of the problem

to be that of the remaining individual hypothesis. An alternative that still gives strong control of the FWER is to use a combination test to find the combined P-values of the individual hypotheses and then apply a closed testing procedure on these combined P-values.

Example

Under the first formulation of the problem, if we apply the alternative overall testing procedure, the pre-interim recruitment cohort contributes $P_1^{(1)}$ and $P_3^{(1)}$. If the trial continues in both the sub-population and the full population then the post-interim recruitment cohort contributes $P_1^{(2)}$ and $P_3^{(2)}$. Otherwise if the trial continues only in the sub-population this only gives $P_1^{(2)}$ and we set $P_3^{(2)} = 1$. We apply the weighted inverse normal combination test to find $P_1^{(c)}$ from $P_1^{(1)}$ and $P_1^{(2)}$ and $P_3^{(c)}$ from $P_3^{(1)}$ and $P_3^{(2)}$; to which we apply Simes method to find the p-value for the intersection hypothesis $P_{13}^{(c)}$ and apply the closed testing procedure.

Similarly under the second formulation of the problem we may apply the alternative testing procedure, the pre-interim recruitment cohort contributes $P_1^{(1)}$ and $P_2^{(1)}$. If the trial continues in both the sub-population and the full population then the post-interim recruitment cohort contributes $P_1^{(2)}$ and $P_2^{(2)}$. Otherwise if the trial continues only in the first sub-population this only gives $P_1^{(2)}$ and we set $P_2^{(2)} = 1$ and if the trial continues only in the second sub-population we have $P_2^{(2)}$ and set $P_1^{(2)} = 1$. Applying the weighted inverse normal combination test we find the combined p-values for the individual null hypotheses $P_1^{(c)}$ from $P_1^{(1)}$ and $P_1^{(2)}$

Formulation 1	
Intersection p-value	$\mathbb{E}_{\pi(\boldsymbol{\theta})}(G(\boldsymbol{\theta}, \kappa))$
Within each stage	3.88
After combination test	3.90
Improved Simes method within each stage	3.94

Formulation 2	
Intersection p-value	$\mathbb{E}_{\pi(\boldsymbol{\theta})}(G(\boldsymbol{\theta}, \kappa))$
Within each stage	2.44
After combination test	2.38

Table 7.7: Comparing $\mathbb{E}_{\pi(\boldsymbol{\theta})}(G(\boldsymbol{\theta}, \kappa))$ using different combination tests under each formulation of the problem

and $P_2^{(c)}$ from $P_2^{(1)}$ and $P_2^{(2)}$; to which we apply Simes method to find the p-value for the intersection hypothesis $P_{12}^{(c)}$ and apply the closed testing procedure.

For each formulation of the problem Table 7.7 compares the Bayes expected gain of the Bayes optimal Adaptive Enrichment trials when finding the intersection p-values within each stage or at the end. We see that changing the testing order of the Adaptive Enrichment design provides an improvement in the Bayes expected gain for formulation 1 but reduces the Bayes expected gain in formulation 2, again showing that there is no clearly better testing method. This shows that the hypothesis testing method should be carefully selected for any particular trial in order to make the best choice.

It is worth noting that under the first formulation of the problem the small improvement offered by changing the testing order is not as large as using the improved version of Simes method within each stage; the improved version of Simes method cannot be used at the end of the trial as the correlation is not known until the interim decision has been made.

7.2 Early stopping for futility

Stopping trials early for futility is a common tool available in clinical trials, there is a dual benefit to this it both saves resources for the trial sponsor when the treatment does not look promising and patients are protected from adverse

side effects of treatments that will not help them. To an extent, by selecting an appropriate sub-population for the post-interim recruitment cohort the Adaptive Enrichment trial achieves early stopping for a proportion of the population. We may extend this to allow early stopping of both populations for futility.

Previously we made general definitions of gain functions for the first and second formulations of the problem in equations 4.5 and 4.6, under the first formulation this is

$$G(\boldsymbol{\theta}, \kappa) = \gamma_1(\boldsymbol{\theta})\mathbb{I}(\mathcal{R}_1) + \gamma_3(\boldsymbol{\theta})\mathbb{I}(\mathcal{R}_3) + \gamma_b(\boldsymbol{\theta})\mathbb{I}(\mathcal{R}_b)$$

and similarly under the second formulation of our problem we use

$$G(\boldsymbol{\theta}, \kappa) = \gamma_1(\boldsymbol{\theta})\mathbb{I}(\mathcal{R}_1) + \gamma_2(\boldsymbol{\theta})\mathbb{I}(\mathcal{R}_2) + \gamma_b(\boldsymbol{\theta})\mathbb{I}(\mathcal{R}_b).$$

Under both formulations of the problem the event \mathcal{R}_b is only possible when continuing in both sub-populations, with the event \mathcal{R}_3 also only being possible when continuing in both sub-populations for the first formulation of the problem, by choosing to enrich the trial into either population the corresponding outcome \mathcal{R}_1 or \mathcal{R}_2 becomes more likely. Each of the possible decisions has a clear motivation from the choice of the gain function, when introducing other possible decisions it should be clear in the gain function where the benefit of these possible decisions comes from.

We must add the possibility of early stopping to the gain function in order to motivate this decision at the interim analysis. Define \mathcal{S} as the event that the trial is stopped at the interim analysis, the gain of stopping at the interim analysis is given by $\gamma_{\mathcal{S}}(\boldsymbol{\theta})$. We define the overall gain function to be

$$G(\boldsymbol{\theta}, \kappa) = \gamma_1(\boldsymbol{\theta})\mathbb{I}(\mathcal{R}_1) + \gamma_2(\boldsymbol{\theta})\mathbb{I}(\mathcal{R}_3) + \gamma_b(\boldsymbol{\theta})\mathbb{I}(\mathcal{R}_b) + \gamma_{\mathcal{S}}(\boldsymbol{\theta})\mathbb{I}(\mathcal{S}). \quad (7.4)$$

under the first formulation of the problem and

$$G(\boldsymbol{\theta}, \kappa) = \gamma_1(\boldsymbol{\theta})\mathbb{I}(\mathcal{R}_1) + \gamma_2(\boldsymbol{\theta})\mathbb{I}(\mathcal{R}_2) + \gamma_b(\boldsymbol{\theta})\mathbb{I}(\mathcal{R}_b) + \gamma_{\mathcal{S}}(\boldsymbol{\theta})\mathbb{I}(\mathcal{S}) \quad (7.5)$$

under the second formulation of the problem.

Example

Returning to our example we set $\gamma_S(\boldsymbol{\theta}) = 1.5$, and our specific choice for the gain function for the first formulation of the problem given in Equation 7.4 becomes

$$G(\boldsymbol{\theta}, \kappa) = \lambda\theta_1\mathbb{I}(\mathcal{R}_1) + (\lambda\theta_1 + (1 - \lambda)\theta_2)\mathbb{I}(\mathcal{R}_3) + (\lambda\theta_1 + (1 - \lambda)\theta_2)\mathbb{I}(\mathcal{R}_b) + 1.5\mathbb{I}(\mathcal{S}),$$

similarly under the second formulation of the problem Equation 7.5 becomes

$$G(\boldsymbol{\theta}, \kappa) = \lambda\theta_1\mathbb{I}(\mathcal{R}_1) + (1 - \lambda)\theta_2\mathbb{I}(\mathcal{R}_2) + (\lambda\theta_1 + (1 - \lambda)\theta_2)\mathbb{I}(\mathcal{R}_b) + 1.5\mathbb{I}(\mathcal{S}),$$

Under these choices of gain function the Bayes optimal decision boundaries remain the same when choosing whether to enrich the trials or not. The trial is stopped early for futility when the estimated treatment effect in both sub-populations is low, this means that there is a minimum value for the Bayes expected gain for the remainder of the trial at the interim analysis. This is similar to stopping the trial early for futility when the observed treatment effect at the interim analysis is below some threshold, but achieved through the gain function that has been central to our optimisation structure. We choose $\gamma_S(\boldsymbol{\theta}) = 1.5$ in this case to illustrate the impact of early stopping in our decision framework, in practice more thought would be required for this value; it may represent the cost of sampling, or perhaps an opportunity cost where the resources used for a trial may be allocated elsewhere if it is stopped early for futility.

Table 7.8 shows the Bayes expected gain of the Bayes optimal Adaptive Enrichment design that allows early stopping for futility with the fixed sampling designs. Under both formulations of the problem we see that stopping early for futility has increased the Bayes expected gain of the Adaptive Enrichment trial. For both formulations of the problem the increase in performance from the early stopping is similar to the increase we see from the Bayes optimal Adaptive Enrichment trial when compared to the fixed sampling design testing multiple hypotheses. We could of course apply this method of early stopping to the fixed sampling designs, introducing an interim analysis purely for this purpose.

Trial design	Formulation 1	
	$\mathbb{E}_{\pi(\boldsymbol{\theta})}(G(\boldsymbol{\theta}, \kappa))$	$\mathbb{P}(\text{Stopping early})$
Adaptive Enrichment early stopping	4.06	0.21
Adaptive Enrichment	3.88	-
Fixed Enrichment	3.77	-
Fixed sample multiple hypotheses	3.74	-

Trial design	Formulation 2	
	$\mathbb{E}_{\pi(\boldsymbol{\theta})}(G(\boldsymbol{\theta}, \kappa))$	$\mathbb{P}(\text{Stopping early})$
Adaptive Enrichment early stopping	2.75	0.35
Adaptive Enrichment	2.44	-
Fixed Enrichment	1.91	-
Fixed sample multiple hypotheses	2.12	-

Table 7.8: Comparing $\mathbb{E}_{\pi(\boldsymbol{\theta})}(G(\boldsymbol{\theta}, \kappa))$ for the Adaptive Enrichment design using early stopping

Chapter 8

Conclusions and some possible extensions

8.1 Conclusions

The first part of our work in Chapter 2 was to set up the hypothesis testing structure we use for Adaptive Enrichment designs. This hypothesis testing structure ensures strong control of the familywise error rate for each of the possible decisions at the interim analysis regardless of the method used to make the decision; we used this hypothesis testing structure throughout our examples. In Chapter 7 we discuss some small changes that could be made to this hypothesis testing structure and see that this may improve the overall performance of some of our methods slightly. In particular these changes are useful when we wish to test null hypothesis in a sub-population and the full population.

We have introduced the use of a Bayes decision framework to optimise the interim decision of Adaptive Enrichment trials. This framework is very flexible allowing us to find the Bayes optimal decision at the interim analysis for any choice of the design parameters, where the Bayes optimal decision rule is the best possible decision rule for the set of parameters. In Chapter 7 we demonstrated this flexibility further by adding an additional decision at the interim analysis, allowing for early stopping for futility. Having developed the necessary framework in Chapter 4 we construct Bayes optimal Adaptive Enrichment designs, the example showed that the Bayes optimal adaptive designs

can offer an improvement over fixed sampling alternatives.

In Chapter 4 we also compared the performance of the Bayes optimal decision rule and optimised rules of a simpler form. While the optimised simple rules did give the Adaptive Enrichment designs a higher overall performance than the fixed sampling alternatives the Bayes optimal decision rules were clearly better. The simple decision rules that we used were chosen to be close to the Bayes optimal decision boundary while being simple to implement, in practice if a simple decision rule is chosen poorly even when optimised it will not give a good overall performance for the Adaptive Enrichment design; there is a further example of this in Chapter 6 where we compare the performance of interim decision rules optimised for different prior distributions.

While developing our optimisation structure we used an immediate response in our examples, meaning that all data are available when an analysis is conducted. This is not a requirement of our optimisation structure. In Chapter 5 we began by using a delayed response, unsurprisingly this reduced the overall performance of the trial as decisions must be made on less data but are not able to change the remaining recruitment of the trial any more than an immediate response. We showed that the loss in overall performance can be recovered by the introduction of a secondary endpoint; we assumed a correlation structure between the primary and secondary endpoint and used this to improve the estimate of treatment effect for the primary endpoint at the interim analysis.

In Chapter 6 we examine how Adaptive Enrichment designs perform across a range of possible scenarios. Using our optimisation framework we see that the Adaptive Enrichment designs provide an improvement in the overall performance when compared to these fixed sampling alternatives in many cases; however Adaptive Enrichment designs are not always the best choice. Using our optimisation framework we suggest comparing the potential performance of an Adaptive Enrichment design with the fixed sampling alternatives for a given situation to find which the most suitable choice is. It is possible to use the simplified computation of the Oracle decision rule, Chapter 4, to give an impression of whether an Adaptive Enrichment design may be useful for a

particular scenario.

Overall using the Bayesian decision framework for the optimisation of the interim decision of Adaptive Enrichment designs has proved to be a powerful tool. We have shown that although we must be sure that an Adaptive Enrichment trial is the correct choice, when it is we can provide the best version of an Adaptive Enrichment design in all settings. When used appropriately these Bayes optimal Adaptive Enrichment designs can provide a large improvement in overall performance when compared to fixed sampling alternatives.

8.2 Extensions

8.2.1 Parameters governing recruitment

We have assumed that λ the proportion of the sample in the first sub-population is equal to the proportion that this represents the sub-population in the overall patient population, we could instead consider oversampling a sub-population particularly one that is rare in the patient population. We have also assumed that λ must be consistent between stages of the trial but we could even consider allowing this to be changed for the post-interim recruitment cohort. Our optimisation framework from Chapter 4 would still allow us to optimise these trials and find their overall performance.

We have assumed that there are only two sub-populations under both formulations of the problem, however it may be possible to identify more sub-populations where the treatment is expected to behave differently. Under the second formulation we assumed that our sub-populations are independent, this assumption may not hold. If for example if each sub-population is identified by the presence of a sub-population specific biomarker some patients may be biomarker positive for multiple sub-populations, this will introduce a correlation between the summary statistics of the patient populations.

8.2.2 Sample size

Throughout our examples we have assumed a fixed overall sample size for the Adaptive Enrichment designs, that is the same as we used for the fixed sampling designs. The Bayes expected gain showed that we could achieve a higher overall performance using an Adaptive Enrichment trial in our examples, if the cost of sampling was particularly high we might aim to reduce the sample size using the Adaptive Enrichment trial such that we achieve the same Bayes expected gain as the fixed sampling designs.

Conversely when we looked at early stopping for futility we could consider this as a cost of sampling. With a cost of sampling defined we could consider adapting the sample size at the interim analysis in addition to selecting a sub-population. If after the interim analysis we ran the rest of the trial as a group sequential trial Turnbull and Jennison (2000) this would give the lowest expected sample size for the remainder of the trial, this would give the lowest cost of sampling in terms of the Bayes expected gain for the trial at the time of the interim analysis.

8.2.3 Niche population

In our null hypotheses we have investigated whether the new treatment was equivalent to or better than the control treatment in either sub-population or the full population depending on the formulation of the problem. In practice when conducting the trial to allow for investigation of the sub-population we may be looking for an effect in a niche population, we might look for a higher improvement over the control treatment, for example we might investigate the null hypothesis $H_{niche} : \theta_1 \leq b$ where b is some higher treatment effect.

Without the formal definition of a niche hypothesis under the first formulation of the problem we may reject the null hypotheses in both the sub-population and the full population. In this instance it may not be clear that the result in the full population is not driven by a very high treatment in the sub-population where we claimed we believed the new treatment may work better. In this scenario we may require further proof that the treatment effect is not only from the sub-population. We may for example seek to show that the treatment effect in the complement of

the sub-population is above some threshold, or perhaps incorporate the treatment effects more carefully in the gain function.

8.2.4 Similar and simplified decisions

In the final example of Section 4.4 we saw that optimising a simple decision rule was able to give an overall Bayes expected gain close to that of the Bayes optimal decision rule. The Bayes optimal decision boundaries that we have computed show exactly where the alternative interim decisions are equivalent in terms of the Bayes expected gain for the remainder of the trial, the simple rules were chosen to be similar to the Bayes optimal decision rules which is why their optimal behaviour is close. We could consider similar decisions at the interim analysis, where the expected gain of two or more decisions is close we could make either decision and still be close to optimal. We could investigate how large a region may be defined as a similar decision without a large impact on the overall Bayes expected gain of the trial. From this we may inform simplified decision rules for the trial, where the simplified rule is contained entirely within the similar region.

8.2.5 Overall treatment effect estimates

When discussing the possible outcomes of the trials we have only considered which null hypotheses have been rejected. In addition to details about hypothesis tests we will also wish to make inferences about the treatment effects after conducting the trial. Constructing the estimates and confidence intervals corresponding to the results of the Adaptive Enrichment trial are not trivial, as using the pooled data will not necessarily correspond with the results of the overall hypothesis testing procedures we have used. These treatment effect estimates could be used to enhance the information passed to the gain function, giving a clearer idea of what gain is achieved through any particular outcome.

8.2.6 Multiple interim analyses

A final area that our optimisation framework could easily be extended to is the idea of conducting multiple interim analyses. In both of the formulations of the problem we have presented there is only one adaptation to be made, and we may

only make this adaptation at the interim analysis. We have seen that changing the timing of the interim analysis can be of benefit for the Adaptive Enrichment trials.

We could offer more opportunities to make this adaptation, for example we might conduct two interim analyses where we may enrich the trial at the second analysis if we have not already done so at the first. If we have not already made the adaptation by the time of the second interim analysis the optimisation at this point is exactly as we saw previously, we evaluate the Bayes expected gain for the rest of the trial for each possible decision and choose the decision that maximises this. At the first interim analysis there is an additional level to the computation, for the decision that continues the trial in both sub-populations at the interim analysis the expected gain also depends on the optimal decision that may be made at the second interim analysis.

This could be extended to allow for any number of interim analyses up to a maximum of conducting an interim analysis after every patient, although this would not be practical. Extending to more interim analyses would mean our optimisation at any particular analysis would have to account for all future decisions which would also need to be optimised, and so the computation time would be expected to increase dramatically with each additional analysis.

Appendix A

A summary of techniques for practical use

A.1 Recruitment of trials

In Section 3.1.2 we introduce the two formulations of the problem that we have worked with throughout. Under the first formulation of the problem we have a single sub-population of interest within a full population. We define θ_1 as the treatment effect in the sub-population and θ_2 as the treatment effect in the complement, the proportion of the sample in the sub-population is given by λ and hence the treatment effect in the full population is given by $\theta_3 = \lambda\theta_1 + (1 - \lambda)\theta_2$. We aim to test the null hypotheses $H_{01} : \theta_1 \leq 0$ in the sub-population and $H_{03} : \theta_3 \leq 0$ in the full population.

Under the second formulation of the problem we have a two separate sub-populations. We define θ_1 as the treatment effect in the first sub-population and θ_2 as the treatment effect in the second, the proportion of the sample in the first sub-population is given by λ as previously. In this formulation of the problem we ignore the full population aiming to test the null hypotheses $H_{01} : \theta_1 \leq 0$ in sub-population 1 and $H_{03} : \theta_3 \leq 0$ in sub-population 2.

In Chapter 3 we made the assumption that the variance would be the same for both the new treatment and the control treatment, σ^2 say, and that our sample size, n say, would be split equally between the new treatment and the control.

Suppose we conduct a trial to test the null hypothesis $H_0 : \theta_0 \leq 0$, then we have that our estimate of the treatment effect follows the distribution given in Equation 2.2

$$\hat{\theta} \sim N\left(\theta, \frac{4\sigma^2}{n}\right).$$

We assume that the ratio of σ^2 and n takes some fixed value, $\tilde{\mathcal{I}}$ say, where

$$\tilde{\mathcal{I}} = \frac{4\sigma^2}{n}.$$

across all of the trials we then write the distributions of the summary statistics in terms of $\tilde{\mathcal{I}}$.

Fixed sampling trials recruitment

Under both formulations of the problem we may conduct a fixed Enrichment trial in the first sub-population to test the null hypothesis $H_{01} : \theta_1 \leq 0$ as introduced in Section 3.3.6. Similarly under the second formulation of the problem we may conduct a fixed Enrichment trial in the first sub-population to test the null hypothesis $H_{02} : \theta_2 \leq 0$. In both cases the summary statistic $\hat{\theta}_i$ for $i = 1, 2$ follows the distribution given in Section 3.6

$$\hat{\theta}_i \sim N(\theta_i, \tilde{\mathcal{I}}^{-1}).$$

The other fixed sampling alternative we consider is sampling from both sub-populations and testing both null hypotheses, this was first discussed in Section 3.3. Under both formulations of the problem we find summary statistics in each of the sub-populations, $\hat{\theta}_1$ and $\hat{\theta}_2$ say, which follow the distributions given in Section 3.6

$$\hat{\theta}_1 \sim N(\theta_1, (\lambda\mathcal{I}(\theta))^{-1})$$

and

$$\hat{\theta}_2 \sim N(\theta_2, ((1 - \lambda)\mathcal{I}(\theta))^{-1}).$$

Under the first formulation of the problem the estimate of the treatment effect in the full population, $\hat{\theta}_3$ is given by $\hat{\theta}_3 = \lambda\hat{\theta}_1 + (1 - \lambda)\hat{\theta}_2$ and follows the joint

distribution given in Section 3.6

$$\begin{pmatrix} \hat{\theta}_1 \\ \hat{\theta}_3 \end{pmatrix} \sim N_2 \left(\begin{pmatrix} \theta_1 \\ \theta_3 \end{pmatrix}, \begin{pmatrix} (\lambda \mathcal{I}(\theta))^{-1} & \mathcal{I}(\theta)^{-1} \\ \mathcal{I}(\theta)^{-1} & \mathcal{I}(\theta)^{-1} \end{pmatrix} \right).$$

Adaptive Enrichment recruitment

The recruitment of the Adaptive Enrichment trials is split between the pre and post-interim recruitment cohorts as explained in Section 3.5.1. In addition to λ being the proportion of the sample in the first sub-population we define τ as the proportion of the sample from the pre-interim recruitment cohort. Under the second formulation of the problem the pre-interim recruitment cohort has estimates of the treatment effect $\hat{\theta}_1^{(1)}$ and $\hat{\theta}_2^{(1)}$ which follow the distributions given in Section 3.6

$$\hat{\theta}_1^{(1)} \sim N(\theta_1, (\lambda \tau \tilde{\mathcal{I}})^{-1})$$

and

$$\hat{\theta}_2^{(1)} \sim N(\theta_2, ((1 - \lambda) \tau \tilde{\mathcal{I}})^{-1}).$$

The post-interim recruitment cohort may then follow one of three options: If the trial continues in both sub-populations then the estimates of the treatment effect from the post-interim recruitment cohort, $\hat{\theta}_1^{(2)}$ and $\hat{\theta}_2^{(2)}$ say, are distributed as given in Section 3.6

$$\hat{\theta}_1^{(2)} \sim N(\theta_1, (\lambda(1 - \tau) \tilde{\mathcal{I}})^{-1})$$

and

$$\hat{\theta}_2^{(2)} \sim N(\theta_2, ((1 - \lambda)(1 - \tau) \tilde{\mathcal{I}})^{-1});$$

if the trial continues in only a single sub-population for $i = 1, 2$ depending on which population is picked the corresponding estimate of the treatment effect is distributed as in Section 3.6,

$$\hat{\theta}_i^{(2)} \sim N(\theta_i, ((1 - \tau) \tilde{\mathcal{I}})^{-1}) \text{ for } i = 1, 2$$

Under the first formulation of the problem also find estimates of the treatment effect in the full population where possible. In the pre-interim recruitment cohort

this is given by $\hat{\theta}_3^{(1)} = \lambda\hat{\theta}_1^{(1)} + (1 - \lambda)\hat{\theta}_2^{(1)}$ where

$$\begin{pmatrix} \hat{\theta}_1^{(1)} \\ \hat{\theta}_3^{(1)} \end{pmatrix} \sim N_2 \left(\begin{pmatrix} \theta_1 \\ \theta_3 \end{pmatrix}, \begin{pmatrix} (\lambda\tau\tilde{\mathcal{I}})^{-1} & (\tau\tilde{\mathcal{I}})^{-1} \\ (\tau\tilde{\mathcal{I}})^{-1} & (\tau\tilde{\mathcal{I}})^{-1} \end{pmatrix} \right).$$

If the trial continues in both sub-populations then the estimate of the treatment effect from the post-interim recruitment cohort is given by $\hat{\theta}_3^{(2)} = \lambda\hat{\theta}_1^{(2)} + (1 - \lambda)\hat{\theta}_2^{(2)}$

$$\begin{pmatrix} \hat{\theta}_1^{(2)} \\ \hat{\theta}_3^{(2)} \end{pmatrix} \sim N_2 \left(\begin{pmatrix} \theta_1 \\ \theta_3 \end{pmatrix}, \begin{pmatrix} (\lambda(1 - \tau)\tilde{\mathcal{I}})^{-1} & ((1 - \tau)\tilde{\mathcal{I}})^{-1} \\ ((1 - \tau)\tilde{\mathcal{I}})^{-1} & ((1 - \tau)\tilde{\mathcal{I}})^{-1} \end{pmatrix} \right).$$

A.2 Testing methodologies

A.2.1 FamilyWise Error Rate

For all trials we have considered throughout this work we have required strong control of the FamilyWise Error Rate (FWER). This was introduced in Section 3.2.1, where strong control is defined in Equation 3.1 as

$$\mathbb{P}_{\boldsymbol{\theta}}(\text{Reject at least one true null hypothesis}) \leq \alpha \text{ for all } \boldsymbol{\theta}.$$

A.2.2 Closed testing procedure

We have used closed testing procedures to ensure strong control of the FWER, we introduce these in Section 3.2.2. Suppose we have two null hypotheses $H_{01} : \theta_1 \leq 0$ and $H_{02} : \theta_2 \leq 0$. In order to reject H_{01} globally at level α the individual tests of H_{01} and $H_{01} \cap H_{02}$ must both be rejected. Similarly in order to reject H_{02} globally at level α the individual tests of H_{02} and $H_{01} \cap H_{02}$ must both be rejected.

In Section 3.2.3 we noted that any testing procedure that gives strong control of the FWER may be written as a closed testing procedure. Given that we require strong control of the FWER for our trial designs this means that choosing to use a closed testing procedure does not restrict the usefulness of our results as this simply covers all testing procedures.

Simes

The closed testing procedure requires a test of the intersection hypotheses, the first method we use for this is Simes method. We introduced this in Section 3.3.3. Suppose we have two null hypotheses $H_{01} : \theta_1 \leq 0$ and $H_{02} : \theta_2 \leq 0$ with corresponding p-values P_1 and P_2 , the intersection p-value P_{12} say, is given by Equation 3.4 which is

$$P_{12} = \min(2\min(P_1, P_2), \max(P_1, P_2)).$$

We used Simes method in our comparisons of Adaptive Enrichment and fixed sampling alternatives throughout Chapters 4, 5 and 6.

Improved Simes and Dunnett

Under the first formulation of our problem where we consider a sub-population within a full population testing the null hypotheses $H_{01} : \theta_1 \leq 0$ and $H_{03} : \theta_3 \leq 0$. Under this formulation of the problem there is a correlation between the treatment effect in the sub-population and the treatment effect in the full population and hence also the corresponding Z-values, say Z_1 is the Z-value in the sub-population and Z_3 is the Z-value in the full population then we know

$$\begin{pmatrix} Z_1 \\ Z_3 \end{pmatrix} \sim \left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \right).$$

This correlation structure makes Simes method conservative, in Section 7.1 we saw how we could remove this conservatism. Suppose P_1 and P_3 are the p-values corresponding to H_{01} and H_{03} then the intersection P-value P_{13} is given by

$$P_{13} = \min(\psi_\rho \min(P_1, P_3), \max(P_1, P_3)),$$

where ψ_ρ is chosen such that Simes method spends the full α . We saw in the example of Section 7.1.1 that the improved Simes method slightly increased the overall performance of both the Adaptive Enrichment trial and the fixed sampling trial testing multiple hypotheses.

An alternative to the improved version of Simes method is our version of the Dunnett type method seen in Section 7.1.2. Let

$$d = \max(z_1, z_3)$$

the P-value for the intersection hypothesis is given by Equation 7.2 which is

$$p_{13} = \mathbb{P}(Z_1 \geq d \cup Z_3 \geq d).$$

As with the improved Simes method this uses the full α available for the hypothesis test. The example of Section 7.1.2 shows that the Dunnett type method improves the overall performance of both the Adaptive Enrichment trial and the fixed sampling trial testing multiple hypotheses. The improved Simes method gave a larger improvement to the fixed sampling trial and the Dunnett method gave a larger improvement for the adaptive trial in this case. Given that there is no universally preferred option we recommend choosing the hypothesis testing methods to suit any particular trial.

A.2.3 Combination tests

In Section 3.4 we introduced the concept of a combination test for the Adaptive Enrichment trial. Combination tests are necessary since the Adaptive Enrichment trial is split into two stages, where the choice of what will be observed in the second stage depends on the first stage observations. For any analysis we split the data according to when the patient was recruited, obtaining an estimate of the treatment effect for the pre-interim and post-interim recruitment cohorts. That is for a treatment effect, θ say, we find $\hat{\theta}^{(1)}$ from the pre-interim recruitment cohort and $\hat{\theta}^{(2)}$ from the post-interim recruitment cohort where $\hat{\theta}^{(1)}$ and $\hat{\theta}^{(2)}$ will be conditionally independent for the Adaptive Enrichment trials.

Weighted inverse normal

The first form of combination test that we have used is the weighted inverse normal that we introduced in Section 3.4.1. Suppose $Z^{(1)}$ and $Z^{(2)}$ are the Z-values corresponding to $\hat{\theta}^{(1)}$ and $\hat{\theta}^{(2)}$. We find the combined Z-value, $Z^{(c)}$ say, for

testing the null hypothesis $H_0 : \theta_0 \leq 0$ at the end of the trial from Equation 3.5,

$$Z^{(c)} = w_1 Z^{(1)} + w_2 Z^{(2)}.$$

Where the weights w_1 and w_2 are such that $\sum_{i=1}^2 w_i^2 = 1$. When these weights are proportional to the sample size in each recruitment cohort this test achieves the highest power.

χ^2 combination test

An alternative the weighted inverse Normal combination test is the χ^2 combination test . Suppose $P^{(1)}$ and $P^{(2)}$ are the P-values corresponding to the Z-values $Z^{(1)}$ and $Z^{(2)}$, then

$$-2\log(P^{(1)}P^{(2)}) \sim \chi_4^2.$$

The combined p-value for testing the null hypothesis at the end of the trial is given by Equation 7.3

$$p^{(c)} = \mathbb{P}(-2\log(P^{(1)}P^{(2)}) \geq -2\log(p^{(1)}, p^{(2)}))$$

This combination test is not as powerful as the weighted inverse Normal when the weights are correct, we are able to choose reasonable weights for Adaptive Enrichment designs so the χ^2 combination test does not seem to be particularly useful.

A.2.4 Overall testing procedure

In Section 3.5.2 we discussed the options for the overall testing procedure of Adaptive Enrichment trials. The first option that we use in most of our examples is to find the P-value for the intersection hypothesis within each stage of the trial and then use a combination test to find the overall p-values. For example under the second formulation of the problem in the pre-interim recruitment cohort we find $P_1^{(1)}$ in the first sub-population and $P_2^{(1)}$ in the second sub-population in the usual way and then find the intersection p-value $P_{12}^{(1)}$ using Simes method.

In the post-interim recruitment cohort if the trial has continued in both sub-populations then we find the p-values $P_1^{(2)}$, $P_2^{(2)}$ and $P_{12}^{(2)}$ in the same way. If the trial has only continued in sub-population i where $i = 1$ or 2 then we set $P_{12}^{(2)} = P_i^{(2)}$. The overall p-values $P_1^{(c)}$, $P_2^{(c)}$ and $P_{12}^{(c)}$ are found using a combination test.

The second option is to find the combined p-values for the individual hypotheses first and then find the intersection p-value. That is we find $P_i^{(c)}$ from $P_i^{(1)}$ and $P_i^{(2)}$ using a combination test for $i = 1, 2$, then from $P_1^{(c)}$ and $P_2^{(c)}$ we find $P_{12}^{(c)}$ using Simes method. Whichever method is used we reject H_{01} globally at level α if $P_1^{(c)} \leq \alpha$ and $P_{12}^{(c)} \leq \alpha$. Similarly we reject H_{02} globally at level α if $P_2^{(c)} \leq \alpha$ and $P_{12}^{(c)} \leq \alpha$ as given by the closed testing procedure.

Both overall testing procedures ensure strong control of the FWER and in our example we do not see a large difference in the overall performance of the Adaptive Enrichment trial whichever is used. It is worth noting that under the first formulation of the problem there is a larger benefit in using either the improved Simes method or Dunnett type method for the intersection hypothesis within each stage of the trial (they cannot be applied at the end as the correlation is not pre-defined).

A.3 Defining the gain function

The key component from our decision framework is to define a single measure of trial performance, we do this by the introduction of gain functions in Section 4.2.2. If we define \mathcal{R}_1 as rejecting the null hypothesis in the first sub-population, \mathcal{R}_2 as rejecting the null hypothesis in the second sub-population, \mathcal{R}_3 as rejecting the null hypothesis in the full population and \mathcal{R}_b as rejecting the both null hypothesis under a particular formulation of the problem. Under the first formulation of the problem we define the general form of the gain function in Equation 4.5

$$G(\boldsymbol{\theta}, \kappa) = \gamma_1(\boldsymbol{\theta})\mathbb{I}(\mathcal{R}_1) + \gamma_3(\boldsymbol{\theta})\mathbb{I}(\mathcal{R}_3) + \gamma_b(\boldsymbol{\theta})\mathbb{I}(\mathcal{R}_b).$$

Under the second formulations of the problem we define the general form of the gain function in Equation 4.6

$$G(\boldsymbol{\theta}, \kappa) = \gamma_1(\boldsymbol{\theta})\mathbb{I}(\mathcal{R}_1) + \gamma_2(\boldsymbol{\theta})\mathbb{I}(\mathcal{R}_2) + \gamma_b(\boldsymbol{\theta})\mathbb{I}(\mathcal{R}_b).$$

The key under both formulations of the problem is that the gain depends on the eventual outcomes of the trial, with the gain of an outcome being given by $\gamma_i(\boldsymbol{\theta})$ for $i = 1, 2, 3, b$. We have allowed the gain of an outcome to depend on the true treatment effects, in practice we could also incorporate the estimates of the treatment effects or other endpoints such as the safety of the new treatment.

A.4 Optimisation of adaptive enrichment

A.4.1 Assessing overall trial performance

In addition to our gain function in Section 4.3.3 we allow for the use of prior distributions to account for our uncertainty about the true treatment effects, say the pdf of this prior distribution is given by $\pi(\boldsymbol{\theta})$. The Bayes expected gain $\mathbb{E}_{\pi(\boldsymbol{\theta})}(G(\boldsymbol{\theta}, \kappa))$ may be used as the measure of overall performance, this is given in Equation 4.24

$$\mathbb{E}_{\pi(\boldsymbol{\theta})}(G(\boldsymbol{\theta}, \kappa)) = \int_{\boldsymbol{\theta} \in \Theta} \pi(\boldsymbol{\theta}) \mathbb{E}_{\boldsymbol{\theta}}(G(\boldsymbol{\theta}, \kappa)) d\boldsymbol{\theta}.$$

Evaluating the Bayes expected gain of different trials for choices of the prior distribution and the gain function will show their comparable overall performance, where the trial that maximises the Bayes expected gain is the best choice.

A.4.2 Optimising simple decision rules

In Section 4.3.4 we saw how we might use this measure of overall performance to optimise decision rules of any particular form. Finding the values of the parameters of the simple decision rules that maximise the Bayes expected gain. Under the first formulation of the problem we optimised a simple decision rule of

the form given in Equation 4.1

$$\text{decision} = \begin{cases} \text{continue in both populations} & \text{if } \hat{\theta}_3^{(1)} \geq \psi \\ \text{continue only in the sub-population} & \text{if } \hat{\theta}_3^{(1)} < \psi. \end{cases}$$

The optimised version of this simple decision rule gave the Adaptive Enrichment trial a similar overall performance to the fixed Enrichment trial. Under the second formulation of the problem we optimised a simple decision rule of the form given in Equation 4.2

$$\text{decision} = \begin{cases} \text{continue in both sub-populations} & \text{if } \hat{\theta}_1^{(1)} \text{ and } \hat{\theta}_2^{(1)} \geq \psi \\ \text{continue only in sub-population 1} & \text{if } \hat{\theta}_1^{(1)} \geq \hat{\theta}_2^{(1)} \text{ and } \hat{\theta}_2^{(1)} < \psi \\ \text{continue only in sub-population 2} & \text{if } \hat{\theta}_2^{(1)} < \hat{\theta}_1^{(1)} \text{ and } \hat{\theta}_1^{(1)} < \psi. \end{cases}$$

The optimised form of this simple decision rule gave an Adaptive Enrichment trial that provided an improvement over all fixed sampling methods that we made comparisons with.

The simple rules that we optimised were chosen knowing the form of the Bayes optimal decision boundary and so are able to perform reasonably well. However applying a simple rule without this knowledge could severely hamper the overall performance of the design; in Section 6.1.4 we see that decision rules of the wrong form meant the Adaptive Enrichment design was no longer the best choice of design in examples where with the right optimisation it provided a clear benefit.

A.4.3 Bayes optimal decisions

Optimising decision rules of any particular form at best limits the overall performance of the trial, in Section 4.4 we introduce the Bayes optimal decision at the interim analysis. Defining κ_1 to be the data available at the time of the interim analysis the Bayes expected gain for the remainder of the trial, $\mathbb{E}_{\pi(\boldsymbol{\theta}), \kappa_1}(G(\boldsymbol{\theta}, \kappa))$ is given by Equation 4.25

$$\mathbb{E}_{\pi(\boldsymbol{\theta}), \kappa_1}(G(\boldsymbol{\theta}, \kappa)) = \int_{\boldsymbol{\theta} \in \Theta} \pi(\boldsymbol{\theta} | \kappa_1) \mathbb{E}_{\boldsymbol{\theta}, \kappa_1}(G(\boldsymbol{\theta}, \kappa)) d\boldsymbol{\theta}.$$

where the choice of sub-populations at the interim analysis that maximises this is the Bayes optimal decision.

We have shown that these Bayes optimal Adaptive Enrichment trials offer an improvement over the fixed sampling alternatives, in Chapter 6 we saw a variety of examples where this was the case. Under the first formulation of the problem we saw that the Bayes optimal Adaptive Enrichment design offered an improvement over the fixed sampling designs in a variety of settings. The advantage of the adaptive design is not typically large, however providing a compromise between fixed sampling designs and not reducing the overall performance of the trial will be beneficial in practice when agreement cannot be achieved over which of the fixed sampling designs should be used.

Under the second formulation of the problem the Bayes optimal Adaptive Enrichment trials do not offer such consistency in out performing the fixed sampling alternatives. If the treatment effects in both sub-populations are similar then the adaptive design offers a larger benefit than it does under the first formulation of the problem, as they mimic the good features of three comparable fixed sampling trials.

However in Section 6.1.2 we tried a prior distribution where the true treatment effect is high in one sub-population, the fixed Enrichment trial in this sub-population was clearly the best choice of trial. As a more general note this shows us that Adaptive Enrichment designs are not always the best choice and we should check whether they are suitable for a particular scenario before using them in practice.

A.4.4 Early stopping for futility

In Section 7.2 we discussed how the gain function will motivate the possible choices at the interim analysis, improving the chances of a possible outcome allowing us a better chance of receiving the corresponding gain. With this in mind we may include any number of possible decisions at the interim analysis so long as they are motivated by the gain function. In Section 7.2 we allow the inclusion of stopping the trial early for futility, defining \mathcal{S} as the event that the

trial is stopped early we write the gain function under the first formulation of the problem is defined by Equation 7.4

$$G(\boldsymbol{\theta}, \kappa) = \gamma_1(\boldsymbol{\theta})\mathbb{I}(\mathcal{R}_1) + \gamma_2(\boldsymbol{\theta})\mathbb{I}(\mathcal{R}_3) + \gamma_b(\boldsymbol{\theta})\mathbb{I}(\mathcal{R}_b) + \gamma_S(\boldsymbol{\theta})\mathbb{I}(\mathcal{S}).$$

Under the second formulation of the problem the gain function giving early stopping is given by Equation 7.5

$$G(\boldsymbol{\theta}, \kappa) = \gamma_1(\boldsymbol{\theta})\mathbb{I}(\mathcal{R}_1) + \gamma_2(\boldsymbol{\theta})\mathbb{I}(\mathcal{R}_2) + \gamma_b(\boldsymbol{\theta})\mathbb{I}(\mathcal{R}_b) + \gamma_S(\boldsymbol{\theta})\mathbb{I}(\mathcal{S}).$$

Under both formulations of the problem this early stopping improved the overall performance of the Adaptive Enrichment trials as should be expected.

A.5 When is Adaptive Enrichment useful

A.5.1 Oracle Adaptive Enrichment

Before we constructed the Bayes optimal decision at the interim analysis we consider the Oracle decision rule in Section 4.3.1. This has the same structure as the Bayes optimal decision but we assume we know the true treatment effects, $\boldsymbol{\theta}$ say. We evaluate the expected gain for the remainder of the trial given $\boldsymbol{\theta}$, this is given by Equation 4.11

$$\mathbb{E}_{\boldsymbol{\theta}, \kappa_1}(G(\boldsymbol{\theta}, \kappa)) = \gamma_1(\boldsymbol{\theta})\mathbb{P}_{\boldsymbol{\theta}, \kappa_1}(\mathcal{R}_1) + \gamma_3(\boldsymbol{\theta})\mathbb{P}_{\boldsymbol{\theta}, \kappa_1}(\mathcal{R}_3) + \gamma_b(\boldsymbol{\theta})\mathbb{P}_{\boldsymbol{\theta}, \kappa_1}(\mathcal{R}_b)$$

under the first formulation of the problem and Equation 4.12

$$\mathbb{E}_{\boldsymbol{\theta}, \kappa_1}(G(\boldsymbol{\theta}, \kappa)) = \gamma_1(\boldsymbol{\theta})\mathbb{P}_{\boldsymbol{\theta}, \kappa_1}(\mathcal{R}_1) + \gamma_2(\boldsymbol{\theta})\mathbb{P}_{\boldsymbol{\theta}, \kappa_1}(\mathcal{R}_2) + \gamma_b(\boldsymbol{\theta})\mathbb{P}_{\boldsymbol{\theta}, \kappa_1}(\mathcal{R}_b).$$

under the second formulation of the problem. As with the Bayes optimal decision rule we choose the sub-population that maximises this. This is less computationally intensive than the Bayes optimal decision rule, and may help us to learn in what cases the Adaptive Enrichment trial will provide a benefit when compared with the fixed sampling alternatives since it gives an upper bound on the possible performance of the Adaptive Enrichment design.

A.6 Additional data types

A.6.1 Delayed response and survival endpoints

Most of our examples have assumed that observations from patients are immediately available, or at least they are available quickly enough that we may use all of the observations from the pre-interim recruitment cohort at the time of the interim analysis. In practice this will not be the case and we are unlikely to wish to halt recruitment while we wait to be able to conduct an interim analysis. In Chapter 5 we remove this assumption of immediate response to see how Adaptive Enrichment performs.

We start in Section 5.1 by assuming a delay between recruiting a patient and observing their response, in fact the analysis of overall performance that we make in the case of delayed response is parallel with assuming that we have a survival endpoint; we show this to be the case in Section 5.3 using log-rank score statistics to summarise the trial. We may still find the Bayes optimal decision at the interim analysis where the Bayes expected gain for the remainder of the trial

$$\mathbb{E}_{\pi(\boldsymbol{\theta}), \kappa_1}(G(\boldsymbol{\theta}, \kappa)) = \int_{\boldsymbol{\theta} \in \Theta} \pi(\boldsymbol{\theta} | \kappa_1) \mathbb{E}_{\boldsymbol{\theta}, \kappa_1}(G(\boldsymbol{\theta}, \kappa)) d\boldsymbol{\theta}$$

now depends on the remaining observations from the pre-interim recruitment cohort in addition to the observations from the post-interim recruitment cohort.

We see that as the proportion of the pre-interim recruitment cohort we have observations from decreases the Bayes expected gain of the Bayes optimal Adaptive Enrichment trials decreases too. This reduction showed a linear trend in our examples, this would allow us to determine a minimum acceptable proportion of observations in order for the trial to provide a benefit. We could also combine this with what we learned in Chapters 6 and 7 in order to optimise all aspects of the trials.

A.6.2 Longitudinal data

When introducing delayed responses in Section 5.1 we stated that one of the reasons for this might be a delay between giving the new treatment and observing the final response, however responses could be collected in a longitudinal fashion with only the final observations contributing to the final analysis. In Section 5.2 we show how using longitudinal observations such as this may be used to enhance the decision at the interim analysis. We compare these optimal decisions with Adaptive Enrichment trials using only the final observations, this shows the improvement from the enhanced decisions. In our examples we recovered approximately half of the performance lost when assuming a delayed response for the observations.

The optimisation methods of Section 5.2 would also apply to survival endpoints under a suitable joint model. We discuss how this may be done in using progression free survival to enhance the interim decision of an Adaptive Enrichment trial where the final analysis is to be conducted based on overall survival in Section 5.3.2, although we do not have a full example of how to simulate the designs in this case.

Bibliography

- P. Bauer and K. Kohne. Evaluation of experiments with adaptive interim analyses. *Biometrics*, pages 1029–1041, 1994.
- J. O. Berger. *Statistical decision theory and Bayesian analysis*. Springer Science & Business Media, 2013.
- D. L. Bhatt and C. Mehta. Adaptive designs for clinical trials. *New England Journal of Medicine*, 375(1):65–74, 2016.
- W. Brannath, M. Posch, and P. Bauer. Recursive combination tests. *Journal of the American Statistical Association*, 97(457):236–244, 2002.
- W. Brannath, E. Zuber, M. Branson, F. Bretz, P. Gallo, M. Posch, and A. Racine-Poon. Confirmatory adaptive designs with bayesian decision tools for a targeted therapy in oncology. *Statistics in medicine*, 28(10):1445–1463, 2009.
- F. Bretz, H. Schmidli, F. Knig, A. Racine, and W. Maurer. Confirmatory seamless phase ii/iii clinical trials with hypotheses selection at interim: General concepts. *Biometrical Journal*, 48(4):623–634, 2006.
- D. Collett. *Modelling survival data in medical research*. CRC press, 2015.
- Committee for Proprietary Medicinal Products and others. Points to consider on multiplicity issues in clinical trials. *London: CPMP*, 2002.
- A. Dmitrienko, R. B. D’Agostino, and M. F. Huque. Key multiplicity issues in clinical drug development. *Statistics in medicine*, 32(7):1079–1111, 2013.
- C. W. Dunnett. A multiple comparison procedure for comparing several treatments with a control. *Journal of the American Statistical Association*, 50(272):1096–1121, 1955.

- A. Elsässer, J. Regnstrom, T. Vetter, F. Koenig, R. J. Hemmings, M. Greco, M. Papaluca-Amati, and M. Posch. Adaptive clinical trial designs for european marketing authorization: a survey of scientific advice letters from the european medicines agency. *Trials*, 15(1):383, 2014.
- R. A. Fisher. *Statistical methods for research workers*. Genesis Publishing Pvt Ltd, 1925.
- Food and Drug Administration and others. Guidance for industry: E9 statistical principles for clinical trials. *Food and Drug Administration: Rockville, Maryland, USA*, 1998.
- Food and Drug Administration and others. Guidance for industry: Adaptive design clinical trials for drugs and biologics. *Washington DC, USA: Food and Drug Administration*, 2010.
- Food and Drug Administration and others. Guidance for industry: Enrichment strategies for clinical trials to support approval of human drugs and biological products. *December. Available at: <http://www.fda.gov/downloads/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/UCM332181.pdf>*, 2012.
- B. Freidlin and R. Simon. Adaptive signature design: an adaptive clinical trial design for generating and prospectively testing a gene expression signature for sensitive patients. *Clinical Cancer Research*, 11(21):7872–7878, 2005.
- K. R. Gabriel. Simultaneous test procedures—some theory of multiple comparisons. *The Annals of Mathematical Statistics*, pages 224–250, 1969.
- A. Genz, F. Bretz, T. Miwa, X. Mi, F. Leisch, F. Scheipl, and T. Hothorn. mvtnorm: Multivariate normal and t distributions. *R package version 0.9-2*, URL <http://CRAN.R-project.org/package=mvtnorm>, 2008.
- H. Götte, M. Donica, and G. Mordenti. Improving probabilities of correct interim decision in population enrichment designs. *Journal of biopharmaceutical statistics*, 25(5):1020–1038, 2015.

- L. V. Hampson and C. Jennison. Group sequential tests for delayed responses (with discussion). *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 75(1):3–54, 2013.
- ICH Steering Committee and others. Statistical principles for clinical trials (e9). In *International Conference on Harmonisation of Technical Requirements for Registration of Pharmaceuticals for Human Use. International Conference on Harmonisation*, 1998.
- International Conference on Harmonisation E9 Expert Working Group and others. Statistical principles for clinical trials: Ich harmonized tripartite guideline. *Statistics in Medicine*, 18(15):1905–1942, 1999.
- S. Irle and H. Schäfer. Interim design modifications in time-to-event studies. *Journal of the American Statistical Association*, 107(497):341–348, 2012.
- M. Jenkins, A. Stone, and C. Jennison. An adaptive seamless phase ii/iii design for oncology trials with subpopulation selection using correlated survival endpoints. *Pharmaceutical statistics*, 10(4):347–356, 2011.
- C. Jennison and B. W. Turnbull. Adaptive seamless designs: selection and prospective testing of hypotheses. *Journal of biopharmaceutical statistics*, 17(6):1135–1161, 2007.
- W. Lehmacher and G. Wassmer. Adaptive sample size calculations in group sequential trials. *Biometrics*, 55(4):1286–1290, 1999.
- R. Marcus, P. Eric, and K. R. Gabriel. On closed testing procedures with special reference to ordered analysis of variance. *Biometrika*, 63(3):655–660, 1976.
- A. O’Hagan, J. W. Stevens, and M. J. Campbell. Assurance in clinical trial design. *Pharmaceutical Statistics*, 4(3):187–201, 2005.
- T. Ondra, S. Jobjörnsson, R. A. Beckman, C.-F. Burman, F. König, N. Stallard, and M. Posch. Optimizing trial designs for targeted therapies. *PloS one*, 11(9):e0163726, 2016.
- M. Posch and P. Bauer. Adaptive budgets in clinical trials. *Statistics in Biopharmaceutical Research*, 5(4):282–292, 2013.

- J. P. Romano, A. Shaikh, and M. Wolf. Consonance and the closure method in multiple testing. *The International Journal of Biostatistics*, 7(1):1–25, 2011.
- M. Rosenblum. Adaptive randomized trial designs that cannot be dominated by any standard design at the same total sample size. 2014.
- S. K. Sarkar. Some probability inequalities for ordered mtp2 random variables: a proof of the simes conjecture. *Annals of Statistics*, pages 494–504, 1998.
- S. K. Sarkar and C.-K. Chang. The simes method for multiple hypothesis testing with positively dependent test statistics. *Journal of the American Statistical Association*, 92(440):1601–1608, 1997.
- H. Schmidli, F. Bretz, A. Racine, and W. Maurer. Confirmatory seamless phase ii/iii clinical trials with hypotheses selection at interim: applications and practical considerations. *Biometrical Journal*, 48(4):635–643, 2006.
- H. Schmidli, F. Bretz, and A. Racine-Poon. Bayesian predictive power for interim adaptation in seamless phase ii/iii trials where the endpoint is survival up to some specified timepoint. *Statistics in medicine*, 26(27):4925–4938, 2007.
- R. J. Simes. An improved bonferroni procedure for multiple tests of significance. *Biometrika*, 73(3):751–754, 1986.
- R. J. Temple. Enrichment designs: efficiency in development of cancer treatments. *Journal of clinical oncology*, 23(22):4838–4839, 2005.
- J. W. Tukey. Quick and dirty methods in statistics. part ii. simple analyses for standard designs. *American Society for Quality Control*, pages 189–197, 1951.
- B. Turnbull and C. Jennison. Group sequential methods with application to clinical trials, 2000.
- US Food and Drug Administration and others. Guidance for the use of bayesian statistics in medical device clinical trials. *Maryland: US Food and Drug Administration*, 2010.
- S.-J. Wang and H. Hung. Adaptive enrichment with subpopulation selection at interim: Methodologies, applications and design considerations. *Contemporary clinical trials*, 36(2):673–681, 2013.

- S.-J. Wang, H. James Hung, and R. T. O'Neill. Adaptive patient enrichment designs in therapeutic trials. *Biometrical Journal*, 51(2):358–374, 2009.
- J. Whitehead. *The design and analysis of sequential clinical trials*. John Wiley & Sons, 1997.